ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

J. Genet. Genomics 35 (2008) 261–266

JOURNAL OF
GENETICS AND
GENOMICS

www.jgenetgenomics.org

Identification of microsatellites in cattle unigenes

Qiuliang Yan [a, b], Yinghan Zhang [a], Hongbin Li [b], Caihong Wei [b], Lili Niu [b], Shan Guan [c], Shangang Li [b], Lixin Du [b, *]

[a] *College of Animal Science and Technology, Northwest A & F University, Yangling 712100, China*
[b] *National Center for Molecular Genetics and Breeding of Animal, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100094, China*
[c] *Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China*

Received for publication 15 August 2007; revised 18 February 2008; accepted 19 February 2008

Abstract

To identify EST-SSR molecular markers, 41,986 cattle UniGene sequences from NCBI were mined for analyzing SSRs. A total of 1,831 SSRs were identified from 1,666 ESTs, which represented an average density of 19.88 kb per SSR. The frequency of EST-SSRs was 4.0%. The dinucleotide repeat motif was the most abundant SSR, accounting for 54%, followed by 22%, 13%, 7% and 4%, respectively, for tri-, hexa-, penta- and tetra-nucleotide repeats. Depending upon the length of the repeat unit, the length of microsatellites varied from 14 to 86 bp. Among the di- and tri-nucleotide repeats, AC/TG (57%) and AGC (12%) were the most abundant type. Annotation of EST-SSRs was also carried out. Three hundred primer pairs were randomly designed using Prime Premier 5.0 program and Oligo 5.0 for further experimental validation.

Keywords: cattle; microsatellite; expressed sequence tags; EST-SSR

Introduction

Microsatellites, also referred to as simple sequence repeats (SSRs), are short repeat motifs (1–6 bp) that are present in both protein coding and non-coding regions of DNA sequences (Gupta et al., 1996; Toth et al., 2000; Katti et al., 2001). Compared to other molecular markers, SSRs are uniquely characterized by their simplicity, abundance, ubiquity, variation, co-dominance and multi-alleles among genomes (Powell et al., 1996). The polymorphism, mainly resulted from the number of repeat units, and can easily be detected by PCR using primers flanking the SSR motif. Microsatellites have become a common tool broadly used in aspects of genetic mapping, molecular evolution and systematic taxonomy in most genomes. However, the development of SSR markers from genomic libraries is expensive, labor intensive and time consuming (Varshney et al., 2002).

Expressed sequence tags (ESTs) are single-pass sequence segments of expressed genes (Adams et al., 1991). Recently, the generation of microsatellite markers using EST sequences has become an attractive alternative to complement existing SSR collections. As a new molecular marker, microsatellite markers derived from EST (EST-SSRs) can be rapidly developed at a low cost. This EST-based approach has been successfully used in plant species such as barley (*Hordeum vulgare* L.) (Thiel et al., 2003), maize (*Zea mays* L.) (Thiel et al., 2003), wheat (*Triticum aestivum* L.) (Hackauf and Lapitan, 2002; Gao et al., 2003; Peng et al., 2005), sugarcane (*Saccharum officinarum* L.) (Pinto et al., 2004) and grape (*Vitis vinifera* L.) (Scott et al., 2000). In animal species, this approach has been used for catfish (*Ictalurus punctatus* A.) (Serapion et al., 2004), shrimp (*Metapenaeus ensis* C.) (Perez et al., 2005) and zebrafish (*Danio rerio* A.) (Ju et al., 2005). In the bovine, Andrés-Mateos et al. (2006) provided an insight into the evolution of the CAG repeat tract at the C-terminus coding region of the *CACNA1A* gene. The objective of this study

* Corresponding author. Fax: +86-10-6281 9997.
E-mail address: lxdu@263.net

was to identify EST-SSR markers and investigate the type and distribution of repeat motifs in the expressed sequence tags of cattle. The results will facilitate the use of molecular markers in cattle breeding.

Materials and methods

Retrieval of UniGene sequences

All cattle ESTs used in this project were directly obtained from NCBI UniGene web-site (ftp://ftp.ncbi.nih.gov/repository/UniGene/) on June 19, 2006. There were 41,986 cattle UniGene clusters containing a total of 1,039,059 ESTs listed and annotated in the database. All the sequences were separately saved in FASTA-Formatted text files that were used as the databases for further analysis.

Detection of SSRs

Cattle UniGene databases were used to identify and characterize SSRs using the Perl program SSRFinder developed for this study (unpublished data). This computer program was run under Linux, using a FASTA-formatted sequence file containing multiple sequences the SSRFinder was written to search all UniGene sequences for all possible combination of di-, tri-, tetra-, penta- and hexa-nucleotide repeats with the criteria of minimum number of 7 repeats for di-nucleotide, 6 for tri-nucleotide, 5 for tetra-nucleotide, 4 for penta-nucleotide and 3 for hexa-nucleotide. Single nucleotide repeats were not selected, because they were generally not considered as useful as polymorphic markers. One file was generated by SSRFinder. The file reported the gene description, number of SSR motifs in each sequence, length and composition of SSR, number of repeats, SSR's start and end position and total length of the EST containing SSR. The frequency of EST-SSRs means the percentage of SSR number identified in cattle UniGene total number.

Annotation of SSR-containing sequences

To obtain an idea about putative functions of SSR-containing genes, these sequences were compared to the nonredundant (nr) protein database of the NCBI Database (http://www.ncbi. nlm.nih.gov/blast) using 1e-05 as a cutoff expected value. The proteins obtained during similarity search were classified into separate groups.

Primer designing for SSRs

For each microsatellite-containing EST, primers were designed using Primer Premier 5.0 and Oligo 5.0 (http://www.co-fly.net). Primer design function was used to determine if the sequences had sufficient flanking sequences for primer design. They were chosen with a length of 17−23 bp, an optimal annealing temperature of 58−63^{o}C and for an amplification product ranging between 150 bp and 400 bp.

Results

Searching for ESTs containing microsatellites

A total of 1,831 SSRs were identified from 1,666 EST sequences, which represented an average density of one SSR/19.88 kb (Table 1). The frequency of EST-SSRs in cattle was 4.0%. The di-inucleotide repeat motifs were the most abundant SSRs in cattle, accounting for 54%, followed by 22%, 13%, 7% and 4% for tri-, hexa-, penta- and tetra-nucleotide repeats. Depending upon the length of the repeat unit itself (2 to 6 bp), the lengths of SSRs varied from 14 to 86 bp. The frequencies of EST-SSRs are presented in Fig. 1.

Frequencies of cattle SSRs with different repeat motifs

Investigation of the distribution of SSR motifs can help gain insights into the composition of genome. The observed frequency of different repeat motifs comprising the SSRs is presented in Figs. 2–6. SSRs identified had 4 types of di-nucleotide repeat motifs, 45 types of tri-nucleotide

Table 1
Summary of *in silico* mining of UniGene sequences of cattle

Parameter	Number
Total number of ESTs	1,039,059
Total number of UniGene sequence searched	41,986
Total number of SSRs identified	1831
Number of sequences containing one SSR	1,522
Number of sequences containing two SSRs	126
Number of sequences containing three SSRs	15
Number of sequences containing four SSRs	3
Total number of SSR-derived EST	1,666

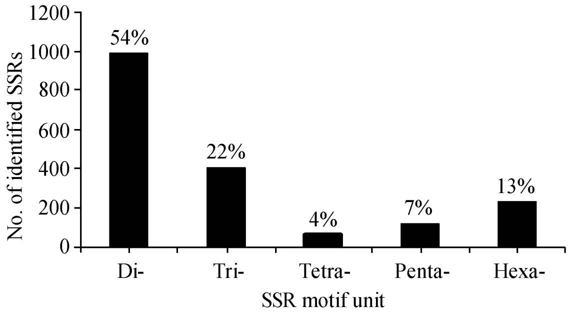

Fig. 1. Frequency distribution of different repeat types (2–6 motif unit) microsatellite identified in UniGene sequences of cattle. The numbers on the bars indicate the percentage of each repeat type microsatellites in total number.

repeat motifs, 49 types of tetra-nucleotide repeat motifs, 48 types of penta-nucleotide repeat motifs and 186 types of hexa-nucleotide repeat motifs.

Among the di-nucleotide repeats, AC/TG was the most abundant type, accounting for 57% of all di-nucleotide repeats found in the cattle ESTs. AT/TA was the second abundant type, accounting for 28%. The GC/CG motif was the least abundant type, only accounting for 1%. The AGC motif was the most abundant tri-nucleotide, followed by GGC and GCG. Tetra-, penta- and hexa-nucleotide repeats with similar frequency and all were AT-rich.

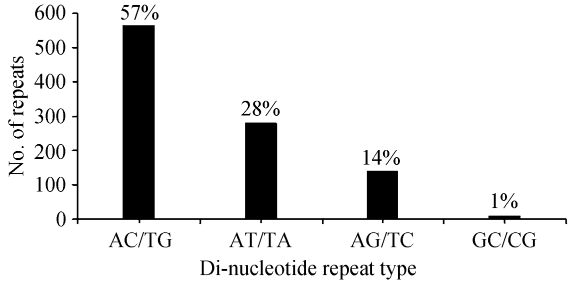

Fig. 2. Frequency distribution of 4 di-nucleotide repeat type in UniGene sequences of cattle. The numbers on the bars indicate the percentage of the 4 di-nucleotide repeat type in all di-nucleotide repeat types.

Annotation of cattle sequences containing SSRs

To determine the function of all the SSR-containing sequences, the 1,666 sequences from which SSRs were mined were annotated against the nonredundant (nr) protein database. BLASTX was used at NCBI (http://www.ncbi.nlm.nih.gov/blast). 698 (41.9%) sequences were annotated genes, and 968 (58.1%) sequences were unknown genes.

Discussion

The results clearly indicate that cattle ESTs are a valuable resource for mining SSR markers. It was found that the frequency of EST-SSRs was 4.0%, with an average of one microsatellite every 19.88 kb of EST sequence.

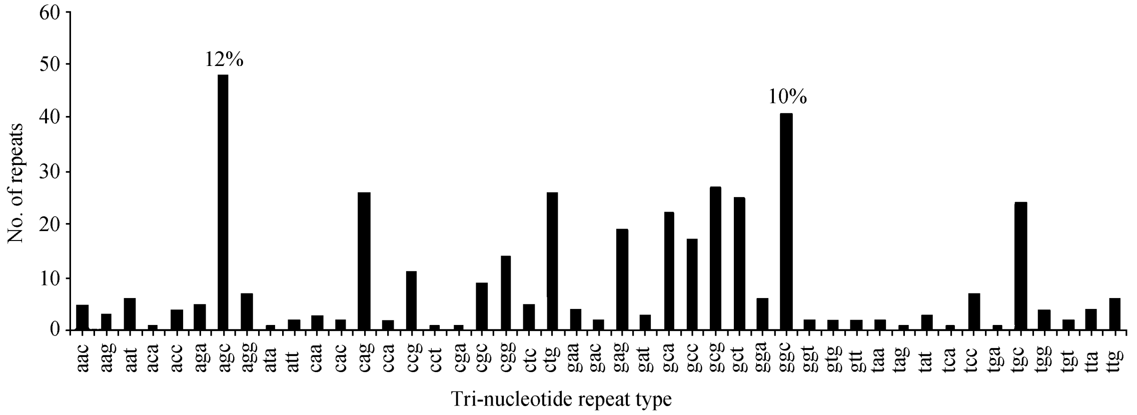

Fig. 3. Frequency distribution of 45 tri-nucleotide repeat type in UniGene sequences of cattle. The percentage of GGC and AGC in all tri-nucleotide repeat type number was lined out on the column.

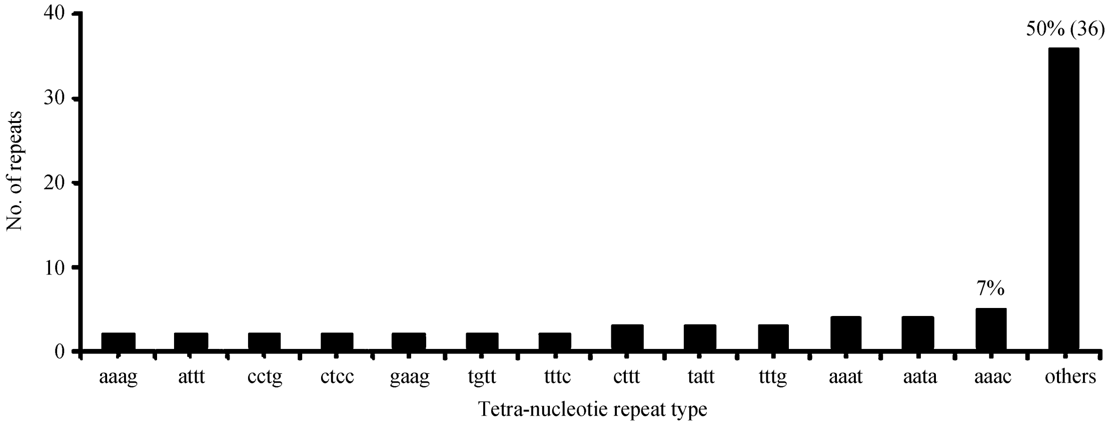

Fig. 4. Frequency distribution of 49 tetra-nucleotide repeat type in UniGene sequences of cattle. Others contain 36 different tetra-nucleotide repeat types and the number of each repeat type is 1, while the percentage of each type is 1.4%. Only the percentage of AAAC and others in all terta-nucleotide repeat type number was lined out on the column.

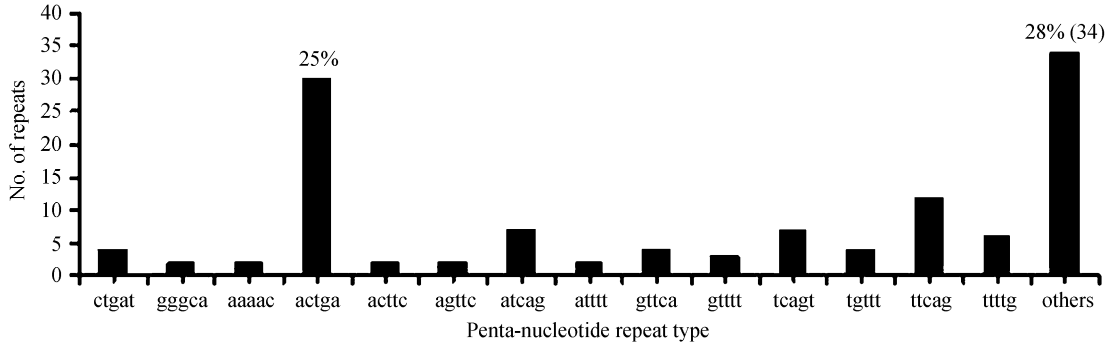

Fig. 5. Frequency distribution of 48 penta-nucleotide repeat type in UniGene sequences of cattle. Others contain 34 different penta-nucleotide repeat types and the number of each repeat type is 1, while the percentage of every type is 0.8%. Only the percentage of ACTGA and others in all penta-nucleotide repeat type number was lined out on the column.

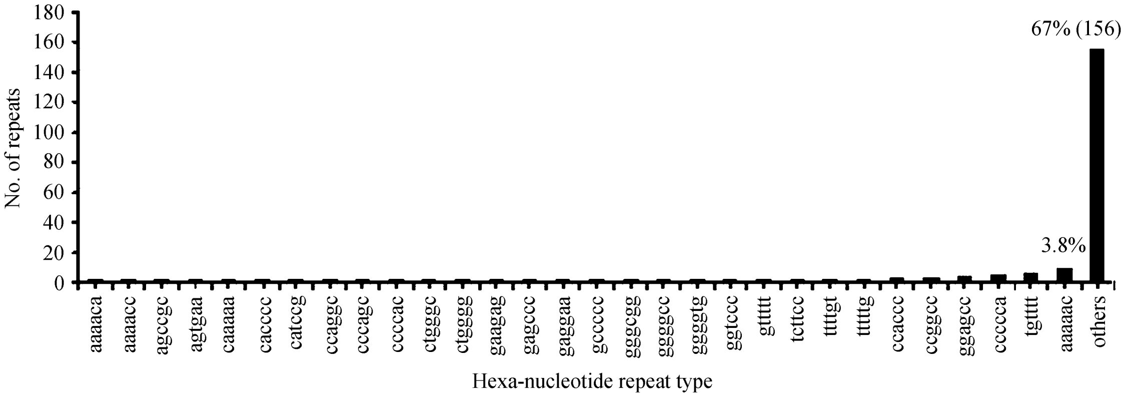

Fig. 6. Frequency distribution of 186 hexa-nucleotide repeat type in UniGene sequences of cattle. Others contain 156 different hexa-nucleotide repeat types and the number of each repeat type is 1, while the percentage of every type is 0.4%. Only the percentage of AAAAAC and others in all hexa-nucleotide repeat type number was lined out on the column.

This EST-SSR frequency was in agreement with that in certain plant species (Kantety et al., 2002; Kumpatla et al., 2005), in which the microsatellite-containing ESTs ranged from 1.5% to 4.7%. The overall frequency and the frequency of different repeat motifs were influenced by redundancy and the criteria used to identify SSR in the mining process.

In this study, di-nucleotide repeats were found to be most abundant, which was in agreement with previous reports on several animal species (Gupta et al., 2002; Rohrer et al., 2002; Serapion et al., 2004; Ju et al., 2005; Perez et al. 2005), but was different from some crop species where tri-nucleotide repeats were abundant (Scott et al., 2000; Cardle et al., 2002; Varshney et al., 2002).

Tri-nucleotide was the second abundant repeat type (22%), followed by hexa- (13%), penta- (7%) and tetra- (4%) nucleotide repeats. The results indicated that the abundance of the different repeats in the SSRs as detected in UniGene sequences was variable. The SSRs with different repeat motifs were thus unevenly distributed. These results were consistent with earlier findings, which showed that the abundance of different repeats varied extensively depending upon the species examined (Toth et al., 2000). The smaller repeat motifs were predominant among SSRs identified. As the length of the repeat unit increases, their occurrence decreases. This can be explained by the fact that longer repeats have higher mutation rates, and therefore are less stable (Wierdl et al., 1997). Furthermore, di-nucleotide and tri-nucleotide repeat stretches were longer than other repeats.

Among the di-nucleotide repeats, AC/TG was the most abundant type, accounting for 57% of all di-nucleotide repeats in the cattle ESTs. AT/TA was the second most abundant type, accounting for 28%, and the GC/CG motif only accounted for 1%. This distribution of di-nucleotide SSRs was similar to what had been found in catfish and zebrafish (Serapion et al., 2004; Ju et al., 2005), but different from that in *Arabidopsis thaliana* (Morgante et al., 2002) and cereal crops (Kantety et al., 2002; Morgante et al., 2002), where AG/TC was the most abundant motif. The AC/TG repeats in plants are less frequent than in animal genomes (Morgante et al., 2002; Toth et al., 2000). This pattern may be related to higher frequencies of certain amino acids in plants than in animals (Toth et al., 2000).

The most abundant tri-nucleotide repeat motif detected in this study was AGC (12%), followed by GGC (10%). These results were in agreement with other reports in the animal kingdom (Li et al., 2004). The distribution of tri-nucleotide motifs in what was different from catfish (Serapion et al., 2004) and zebrafish (Ju et al., 2005) in which the AAT/TAA motif was the most abundant and the tri-nucleotide motifs consisting of only G and /or C or G/C combination were rare, but similar to the plant species in which CCG/GGC was the most abundant type (Nicot et al., 2004; Peng et al., 2005).

Tetra-, penta- and hexa-nucleotide repeat motifs were AT-rich (Figs. 4–6). The distribution of SSR motifs indicated that AT-rich repeats are generally abundant and GC-rich repeats are rare in the transcripts of cattle. Similar observation on the distribution of SSR motifs has been reported in catfish and zebrafish (Serapion et al., 2004; Ju et al., 2005). The presence of SSRs in coding regions shows a bias to some specific nucleotide composition. Olivero et al. (2003) reported that A/T repeats were more frequent than G/C repeats in human coding sequences.

Primers were designed successfully for 300 EST-SSRs. However, the sequences flanking either ends of the SSRs were inadequate in size in many cases. Primer efficiencies remain to be experimentally validated. In this study, only 712 (42.7%) of the 1,666 SSRs containing sequences can be identified, and the remaining 954 (57.3%) were unknown genes.

UniGene sequences of cattle were systematically searched for SSRs using the “SSRFinder” Perl program. EST-SSRs development by data mining has various advantages over conventional development of genomic microsatellites. First, the cost of data mining for EST-SSRs is very low because it avoids the expensive work associated with the initial steps of microsatellite development. Second, as EST-SSR markers are derived directly from expressed genes, product identity and function can be identified by comparing with protein databases, generating type I markers. It can be concluded that data mining can generate abundant EST-SSR markers for a variety of genetic tasks. The identified SSRs would be useful for the development of SSR markers, which are useful in genetic diversity studies and reveal variation in genomes. Annotation of SSR-containing sequences provides an opportunity to examine the functional diversity of different proteins.

Acknowledgements

This work was supported by the National High Science and Technology Foundation of China (863 Project) (No. 2006AA10Z199), Eleventh Five-year Plan on Dairy Industry Key Problem (No. 2006BAD04A01), Animal and Plant Breeding Project of Agriculture and Forestry (No. 2006BAD01A11) and Program of Innovation Fund of IAS.

References

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., and Moreno, R.F. (1991). Complementary DNA sequencing, expressed sequence tags and human genome project. Science **252:** 1651–1656.

Andrés-Mateos, E., Cruces, J., Renart, J., Solís-Garrido, L.M., Serantes, R., de Lucas-Cerrillo, A.M., and Montiel, C. (2006). Bovine *CACNA1A* gene and comparative analysis of the CAG repeats associated to human spinocerebellar ataxia type-6. Gene **380:** 54–61.

Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D., and Waugh, R. (2002). Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics **156:** 847–854.

Gao, L., Tang, J., Li, H., and Jia, J. (2003). Analysis of microsatellites in major crops assessed by computational and experimental approaches. J. Mol. Biol. **12:** 245–261.

Gupta, P.K., Balyan, H.S., Sharma, P.C., and Ramesh, B. (1996). Microsatellites in plants, a new class of molecular markers. Curr. Sci. **70:** 45–54.

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

Hackauf, B., and Wehling, P. (2002). Identification of microsatellite polymorphisms in an expressed portion of the rye genome. Plant Breed. **121:** 17–25.

Ju, Z., Wells, M.C., Martinez, A., Hazlewood, L., and Walter, R.B. (2005). An *in silico* mining for simple sequence repeats from expressed sequence tags of zebrafish, medaka, Fundulus and Xiphophorus. In Silico Biol. **5:** 439–463.

Kantety, R.V., La, R.M., 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.

Katti, M.V., Ranjekar, P.K., and Gupta, V.S. (2001). Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. **18:** 1161–1167.

Kumpatla, S.P., and Mukhopadhyay, S. (2005). Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome **48:** 985–998.

Li, Y.C., Korol, A.B., Fahima, T., and Nevo, E. (2004). Microsatellites within genes, structure, function, and evolution. Mol. Biol. Evol. **21:** 991–1007.

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

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

Olivero, M., Ruggiero, T., Coltella, N., Maffe, A., Calogero, R., Medico, E., and Di Renzo, M.F. (2003). Amplification of repeat-containing transcribed sequences (ARTS), a transcriptome fingerprinting strategy to detect functionally relevant microsatellite mutations in cancer. Nucleic Acids Res. **31:** e33.

Peng, J.H., and Lapitan, N.L. (2005). Characterization of EST-derived microsatellites in the wheat genome and development of eSSR markers. Funct. Integr. Genomics **5:** 80–96.

Perez, F., Ortiz, J., Zhinaula, M., Gonzabay, C., Calderon, J., and Volckaert, F.A. (2005). Development of EST-SSR markers by data mining in three species of shrimp, *Litopenaeus vannamei*, *Litopenaeus stylirostris* and *Trachypenaeus birdy*. Mar. Biotechnol. (N.Y.) **7:** 554–569.

Pinto, L.R., Oliveira, K.M., Ulian, E.C., Garcia, A.A., and de Souza, A.P.(2004). Survey in the sugarcane expressed sequence tag database

(SUCEST) for simple sequence repeats. Genome **47:** 795−804.

Powell, W., Machray, G.C., and Provan, J. (1996). Polymorphism revealed by simple sequence repeats. Trends Plant Sci. **1:** 215−222.

Rohrer, G.A., Fahrenkrug, S.C., Nonneman, D., Tao, N., and Warren, W.C. (2002). Mapping microsatellite markers Identified in porcine EST sequences. Anim. Genet. **33:** 372−376.

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:** 723−726.

Serapion, J., Kucuktas, H., Feng, J., and Liu, Z. (2004). Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). J. Mar. Biotechnol. **6:** 364−377.

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.

Toth, G., Gaspari, Z., and Jurka, J. (2000). Microsatellites in different eukaryotic genome, survey and analysis. Genome Res. **10:** 1967−1981.

Varshney, R.K, Thiel, T., Stein, N., Langridge, P., and Graner, A. (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol. Biol. Lett. **7:** 537−546.

Wierdl, M., Dominska, M., and Petes, T.D. (1997). Microsatellite instability in yeast, dependence on the length of the microsatellite. Genetics **146:** 769−779.