

Received 3 March 2002
Accepted 20 May 2002

**IN SILICO ANALYSIS ON FREQUENCY AND DISTRIBUTION OF
MICROSATELLITES IN ESTs OF SOME CEREAL SPECIES**

RAJEEV K. VARSHNEY¹, THOMAS THIEL¹, NILS STEIN¹,
PETER LANGRIDGE² and ANDREAS GRANER^{1,*}

¹Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3,
D 06466 Gatersleben, Germany; ²Department of Plant Science, University of
Adelaide, Waite Campus, PMB1 Glen Osmond, 5064 South Australia

Abstract: During the last decade microsatellites or SSRs (simple sequence repeats) have been proven to be the markers of choice in plant genetics research and for breeding purposes because of their hypervariability and ease of detection. However, development of these markers is expensive, labour intensive and time consuming, in particular, if they are being developed from genomic libraries. In the context of large-scale sequencing and genomics programmes in various cereal species at different laboratories, a large set of expressed sequence tags (ESTs) is being generated, which can be used to search for microsatellites. Keeping in view the importance of such type of SSRs, available ESTs of some cereal species like barley, maize, oats, rice, rye and wheat were investigated for a study of abundance, frequency and distribution of various types of microsatellites. SSRs were present in about 7% to 10% of the total ESTs in the investigated cereal genomes. On the basis of surveying EST sequences amounting to 75.2 Mb in barley, 54.7 Mb in maize, 43.9 Mb in rice, 3.7 Mb in rye, 41.6 Mb in sorghum and 37.5 Mb in wheat, the frequency of SSRs was 1/7.5 kb in barley, 1/7.5 kb in maize, 1/6.2 kb in wheat, 1/5.5 kb in rye and sorghum and 1/3.9 kb in rice. The overall average SSR frequency for these species is 1/6.0 kb. Trimeric repeats are the most abundant (54% to 78%) class of microsatellites followed by dimeric repeats (17% to 40%). Among the trimeric repeats the motifs CCG are the most common in all the cases ranging from 32% in wheat to 49% in sorghum. When all these SSRs were analysed for assessing their potential to develop new markers, unique primer pairs could be designed for 30% to 70% of the total non-redundant microsatellites which are up to 3% of total ESTs in the studied species.

Key Words: Cereal, EST, Marker, Microsatellite, SSR

* Corresponding author, E-mail: graner@ipk-gatersleben.de

INTRODUCTION

Microsatellites, or simple sequence repeats (SSRs), are stretches of DNA, consisting of tandemly repeated short units of 1-6 base pairs in length. They are ubiquitous in eukaryotic genomes and can be analysed through PCR technology. The sequences flanking specific microsatellite loci in a genome are believed to be conserved within a particular species, across species within a genus and rarely even across related genera. These flanking sequences, therefore, have been used to design primers for individual microsatellite loci and the technique is described as Sequence Tagged Microsatellite Site (STMS) analysis [1] or as Simple Sequence Length Polymorphism (SSLP) [2]. The STMS or SSR markers reveal polymorphisms due to variation in the lengths of microsatellites at specific individual loci; they are therefore, polyallelic and co-dominant in nature, thus proving to be very informative. Consequently, they have been used extensively not only for mapping SSR loci in many cereal species, but also for tagging a number of genes and for diversity studies in these crops [3].

However, development of SSR markers is expensive, labour intensive and time consuming, in particular, if they are being developed from genomic libraries. Even though, due to importance of microsatellites, they have been developed in a large number of plants including major cereal species such as barley [4], maize [5], oats [6], rice [7, 8], rye [9], sorghum [10] and wheat [11, 12]. Due to current emphasis on functional genomics, ESTs (expressed sequence tags) are fast accumulating in EST databases of a large number of crop species. These EST databases can be mined for SSR containing ESTs that would serve for designing locus specific primers. Following this procedure, SSR markers can be obtained at significantly reduced costs, as EST derived SSRs are free by-product of the currently expanding EST databases. While EST-derived SSRs have been shown to be less polymorphic than those derived from genomic sequences [13], they have some intrinsic advantages: they are quickly obtained by electronic sorting, are unbiased in their repeat type, are present in gene rich regions of the genome, and are still abundant [14]. Since they represent the transcribed part of the genome, EST-based SSR markers lead to the direct mapping of genes. Such types of markers are being used presently in only a few crops, as these markers are accessible only in those species for which a sufficient number of ESTs exist in public databases. For instance, the utilization of such type of markers has already been demonstrated in grape [15]. They have been successfully used for genetic diversity studies in wheat [16] and pedigree analysis in barley [17]. Further, compared to SSR markers derived from genomic DNA sequences those based on ESTs have a higher level of transferability among related species as they are located in more conserved regions of the genome. This higher transferability has been demonstrated in grape [15] and sugarcane [18] at different taxonomic levels. For instance, when the ESTs/ genomic DNA derived SSR markers from sugarcane (*Saccharum* spp.) were used to related genera such as *Erianthus* (*Erianthus* spp.) and sorghum (*Sorghum* spp.), EST-derived SSRs

were found to be more superior in terms of transferability [18]. Due to this attractive feature, the EST-derived SSR markers from hexaploid wheat have been successfully used for genotyping the A and B genomes of wheat [19].

In the present study, EST databases of seven major cereal species were mined for the presence of SSRs and assessed their potential for the development of markers. In this context, an *in silico* analysis on the frequency and distribution of microsatellites, and their possible use for the development of genetic markers are presented.

MATERIAL AND METHODS

Sequence data sources

The EST sequences available in the public domain for different cereal species namely barley (*Hordeum vulgare*), maize (*Zea mays*), oats (*Avena sativa*), rye (*Secale cereale*), rice (*Oryza sativa*), sorghum (*Sorghum bicolor*) and wheat (*Triticum aestivum*) were acquired through a Sequence Retrieval System (SRS version 6.1.3 release 28 December, 2001) search of the EMBL database in January 2002. In addition, we also used data from the barley EST-database (B-EST) of the IPK (<http://pgrc.ipk-gatersleben.de/>) presently containing 41,600 ESTs from a variety of different tissues.

Searching of microsatellites

EST sequences less than 100 bp in length were not included in our analysis. Long sequences were shortened at their 3' end to 700 bp due to an expected drop in sequence quality at the end of these sequence reads. The identification and localization of microsatellites was carried out by a Perl5 script (named as *MicroSATellite*, MISA; Thiel *et al.*, unpublished results; the MISA script is available from the IPK website <http://pgrc.ipk-gatersleben.de/misa>), which is capable to identify both perfect microsatellites as well as compound microsatellites (interrupted by a certain number of bases). While classifying the microsatellites into different repeat types or category, sequence complementary was also considered eg. repeat motifs AG, GA, TC and CT were put in the same class.

For searching SSRs by the *MISA* script, microsatellites were considered to contain motifs that are between 1 and 6 nucleotides of size. Therefore a microsatellite was defined as being dinucleotide units repeated 6 times and trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide units repeated 5 times. Once SSR containing ESTs were identified, flanking primers were designed using PRIMER3 software (Whitehead Institute, USA) in a batch modus manner with the help of Perl5 interface modules. To force the picking of flanking primers, the TARGET option was used representing the position of the respective microsatellite enlarged by 3 positions at each side. All other options were left on default value.

For selecting a set of non-redundant SSRs/ primer pairs, repetitive sequences were masked and the Stack_Pack clustering system [20] was used in order to develop unique markers so that non-redundant genes (or unigenes) may be mapped.

RESULTS AND DISCUSSION

Occurrence and frequency of microsatellites

A large set of EST data representing 55.1 Mb in barley, 54.7 Mb in maize, 43.9 Mb in rice, 3.7 Mb in rye, 41.6 Mb in sorghum and 37.5 Mb in wheat was procured from the EMBL database. These EST sequences along with the ESTs of IPK (representing 21.3 Mb) were used for mining microsatellites. The details on mining and analysis of the microsatellites are given in Table 1. We found microsatellites in a range of 6-11% of the total EST sequences from all the cereal species studied except for oats (0.4%) where only a very small dataset was available for analysis. In specific, the frequency of SSRs amounted to 1/7.5 kb in barley, 1/7.5 kb in maize, 1/3.9 kb in rice, 1/5.5 kb in rye and sorghum and 1/6.2 kb in wheat. While calculating the frequency of SSRs, mononucleotides were not taken into account. The estimate of total SSR frequency calculated here indicates the similarity across cereal species and suggests that SSRs occur at the frequency of every 4-8 kb in cereal species. The overall average of SSR frequency for these species is 1/6.0 kb resulting in 42,851 SSRs in a total of 258 Mb of sequence. The highest frequency of one SSR every 3.9 kb was observed in rice which is in close agreement with earlier studies [21, 22]. The frequencies of SSRs in the other species are comparable to earlier results [21] with slight differences. This difference may be explained due to a variation in the quantity of sequence data analysed and differences in defining the criteria for SSR mining in the EST databases.

However, above-mentioned SSR frequencies are based on a redundant set of ESTs. Thus it does not provide a true picture on the frequency of SSRs in the expressed portion of the genome. In the barley genome, for instance, we have demonstrated that the frequency of non-redundant SSRs is 1/ 17.2 kb in the expressed portion of the barley genome on the basis of the identification of 1240 non-redundant SSRs in a set of 41600 ESTs (representing 21.6 MB). The observed 2.2-fold redundancy (17.2 kb/7.5 kb) is expected to increase with the growth of the individual EST databases and needs to be taken into account before the development of markers for a given genome.

Distribution of microsatellite classes

The proportion of the SSR unit size was not evenly distributed in all the cereal species. The trinucleotide repeats, in the range of 54% to 78%, are the most abundant class of microsatellites in the EST sequences of all the species (Figure 1). The dinucleotide, tetranucleotide and pentanucleotide repeats are represented in decreasing proportions of 17.1-40.4%, 2.6-6.6%, and 0.4-1.3%, respectively.

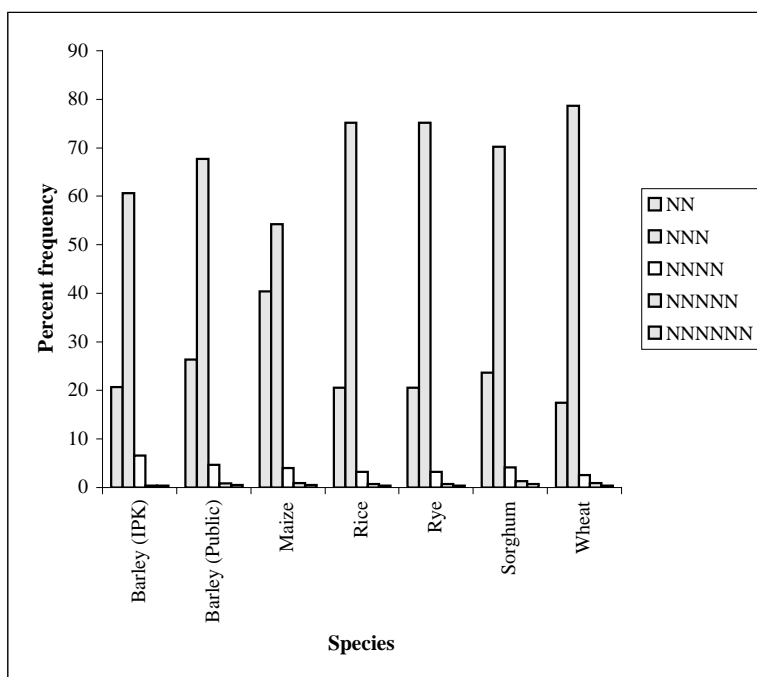


Fig. 1. Distribution of microsatellites (excluding monomeric repeats) in cereal genomes. In all species trimeric repeats are most frequent followed by dimeric and tetrameric while pentameric and hexameric repeats never exceed 1% of the total number of microsatellites.

The hexanucleotide repeats were least frequent (always <1%). These findings are in consistency with previous observations about differences in abundance of SSR unit size classes [21]. In total, it can be concluded that the trimeric SSRs are highly abundant in the EST sequences. This dominance of trimeric SSRs over di-, tetra-, and pentameric SSRs may be explained on the basis of the suppression of non-trimeric SSRs in coding regions due to the risk of frameshift mutations that may occur when those microsatellites alternate in size of one unit [23].

For all species and every class of microsatellites (i.e. di-, tri-, tetra-, penta- and hexameric) the frequency of microsatellites decreases with increasing repeat length (data not shown). In maize, for instance, the single category of SSRs consisting of six repeat units represents 56.9% of the total number of dimeric SSRs and among the trimeric SSRs the category with five repeat units shares as much as 64.5% of the total class (Figure 2). If all microsatellites of different types are classified into two categories of <10 and >10 repeat units, we observe that the category of >10 repeat units contributes only as much as 25% to the total

number of microsatellites (Table 2). In a few cases, especially in the tetrameric, pentameric and hexameric microsatellites all the microsatellites (100%) fall into the category of <10 repeat units.

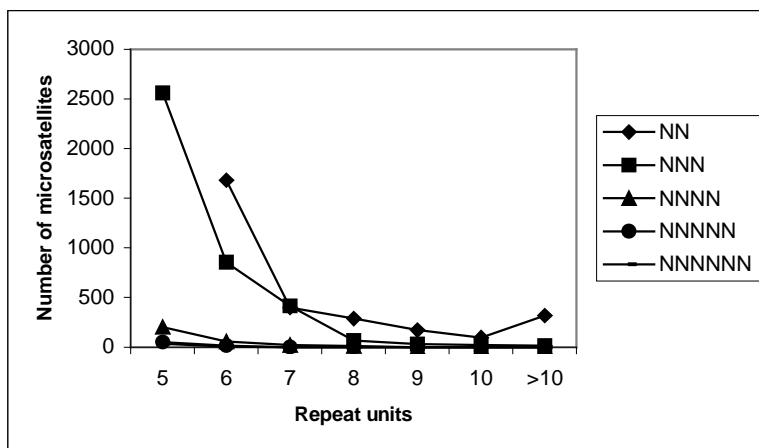


Fig. 2. An example of the distribution of microsatellites of different repeat units in ESTs of maize. The number of microsatellites in a particular class (like NN, NNN, etc.) decreases with the increasing number of repeat units.

Among the dimeric repeats, the motifs AG are the most common (38% to 59%) followed by the motifs AC (20% to 34%) in all the species except of rye. In rye these frequencies are 50% for AC and 37.9% for AG, respectively. The least common motifs are CG in all species (1.7% to 9.0%) except in barley where AT are least common with 8.4%. Overall, the results in the present study are in contrast to earlier studies where the motifs AT were reported to be dominant in plant genomes [8, 24, 25]. However, in some other studies the deficiency of CG and AT motifs has been reported in *Arabidopsis* [21], rice [7] and maize [26]. Our results are based on EST-derived SSRs, they may differ with those SSRs derived from genomic DNA as there might be a systematic bias in the repeat motifs between these two categories.

Among the trimeric repeats, the motifs CCG are the most common in all the cases ranging from 32% in wheat to 49% in sorghum. The abundance of CCG repeats is a specific feature of monocot genomes, which may be due to their increased G+C content [27]. Second to this are AGC (13% to 30%) in barley, maize, rice and sorghum and AAC in rye (16%) and wheat (27%), respectively. The third most frequent motifs in barley, rice, rye and sorghum are AGG while in wheat and maize, they are AGC and AAC, respectively. The abundance of particular trimeric motifs has also been reported for rice [7], maize [26] and sugarcane [18]. Regarding the monocot species investigated for SSRs in ESTs in this study, CCG motifs were the most abundant and AAT motifs the least

common (<1%). This may be explained by the fact that TAA-based variants code for stop codons that have a direct effect on protein synthesis [26]. However, among the tetrameric, pentameric or hexameric microsatellites no specific trend was observed in relation to the abundance of a particular motif class.

For the IPK barley ESTs, we compared the distribution and abundance of a particular class/ repeat motifs of microsatellites in redundant and non-redundant SSRs (Table 3). In both cases, the distribution of different types of microsatellites is comparable. The trimeric repeats, for instance, is the most common class of microsatellites and followed by the dimeric and tetrameric ones. Also the abundance of the AG, CCG and ACGT repeat motifs in the dimeric, trimeric and tetrameric SSRs is similar in both cases. The detailed analysis suggests that, apart from minor deviations, there is no significant difference in the distribution and abundance of microsatellites in the redundant and non-redundant set of barley ESTs.

Tab. 3. Comparative analysis of redundant and non-redundant SSRs in barley ESTs of IPK

	Non-redundant SSRs ¹	Redundant SSRs ²
Total number of SSRs	1240 (3%)	3154 (7.6%)
Distribution of microsatellites (%)		
Dimeric	314 (25.3%)	744 (23.6%)
Trimeric	824 (66.4%)	2183 (69.2%)
Tetrameric	79 (6.3%)	195 (6.2%)
Pentameric	12 (1%)	16 (0.5%)
Hexameric	11 (1%)	16 (0.5%)
Abundance of repeat type		
Dimeric	AG (54.1%)	AG (44.9%)
Trimeric	CCG (39.3%)	CCG (44.8%)
Tetrameric	ACGT (12.7%)	ACGT (19.5%)

¹SSRs identified in non-redundant set of ESTs, selected after clustering; ²SSRs identified in total set of ESTs (41600).

Marker development

For the development of microsatellite markers, we designed primer pairs for all the identified microsatellites. Thus primer pairs could be designed for 53% to 71% of the total microsatellites in the analysed cereal species (Table 1). For the remaining microsatellites, primer pairs could not be designed for one of the following reasons: (a) EST sequences containing microsatellites are too short, (b) microsatellites are too close to the cloning site of the EST, or (c) the flanking sequences are not unique. Thus in total up to 7% of the total EST sequences have a potential for the design of primer pairs in the datasets analysed. These figures are in agreement with earlier studies on EST-derived SSRs in sugarcane [18]. To exclude the development of multiple markers for the same gene, unique primer pairs were identified for non-redundant SSRs by cluster analysis of the corresponding ESTs. Thus, up to 3% of EST sequences can be expected to yield informative SSR markers in the datasets analyzed (Table 1).

CONCLUSION

Using the *MISA*-software tool EST databases can be systematically searched for SSRs for the development of microsatellite markers, which are associated with transcribed genes. This approach saves both costs and time, given a sufficient amount of available EST sequences. However, EST-derived SSR markers are generally less polymorphic in comparison to genomic SSRs as the former are situated in more conserved regions of the genome. Nevertheless, this demerit may be compensated by their higher potential of transferability to related species sometimes even at different taxonomic levels [13, 18]. The use of EST-derived SSRs in genetic diversity studies is a novel tool that reveals variation in transcribed genes. Finally, the *in silico* identification of homoeologous EST sequences in related grass genomes will provide a powerful approach to study marker synteny [28].

Acknowledgements. We thank Dr. Hangning Zhang for trimming, cleaning and managing the barley EST sequences of IPK. This work was supported by a grant from Grains Research and Development Corporation (GRDC), Australia.

REFERENCES

1. Beckmann, J.S. and Soller, M. Towards a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. **Bio/Technology** 8 (1990) 930-932.
2. Tautz, D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. **Nucl. Acids Res.** 17 (1989) 6443-6471.
3. Gupta, P.K. and Varshney, R.K. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. **Euphytica** 113 (2000) 163-185.

4. Ramsay, L., Macaulay, M., Ivanissevich, D.S., MacLean, K., Cardle, L., Fuller, J., Edwards, K.J., Tuveesson, S., Morgante, M., Massari, A., Maestri, E., Marmiroli, N., Sjakste, T., Ganal, M., Powell, W. and Waugh, R.A. A simple sequence repeat-based linkage map of barley. **Genetics** 156 (2000) 1997-2005.
5. Yu, J., Lu, H. and Bernardo, R. Inconsistency between SSR groupings and genetic backgrounds of white corn inbreds. **Maydica** 46 (2001) 133-139.
6. Li, C.D., Rosnagel, B.G. and Scoles, G.J. The development of oats microsatellite markers and their use in identifying relationships among *Avena* species and oats cultivars. **Theor. Appl. Genet.** 101 (2000) 1259-1268.
7. Temnykh, S., Park, W.D., Ayres, N., Cartinhour, S., Hauck, N., Lipovich, L., Cho, Y.G., Ishii, T. and McCouch, S.R. Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). **Theor. Appl. Genet.** 100 (2000) 697-712.
8. Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S. and McCouch, S. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. **Genome Res.** 11 (2001) 1441-1452.
9. Saal, B. and Wricke, G. Development of simple sequence repeat markers in rye (*Secale cereale* L.). **Genome** 42 (1999) 964-972.
10. Bhatramakki, D., Dong, J.M., Chhabra, A.K. and Hart, G.E. An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. **Genome** 43 (2000) 988-1002.
11. Roder, M.S., Korzn, V., Wendehake, K., Plaschke, J., Tixier, M., Leroy, P. and Ganal, M. A microsatellite map of wheat. **Genetics** 149 (1998) 2007-2023.
12. Varshney, R.K., Kumar, A., Balyan, H.S., Roy, J.K., Prasad, M. and Gupta, P.K. Characterization of microsatellites and development of chromosome specific STMS markers in bread wheat. **Plant Mol. Biol. Rep.** 18 (2000) 1-12.
13. Thiel, T. **Identifizierung, Kartierung und Charakterisierung cDNA basierter Mikrosatelliten-Marker zur Diversitätsanalyse bei gerste (*Hordeum vulgare* L.). Diploma thesis.** Technische Universität Dresden (2001) 99.
14. Scott, K.D. Microsatellite derived from ESTs, and their comparison with those derived by other methods. In: **Plant Genotyping: The DNA Fingerprinting of Plants**, (Henry, R. J., Ed.), CABI Publishing, Oxon, U. K., 2001, 225-237.
15. Scott, K D., Eggle, P., Seaton, G., Rossetto, M., Ablett, E.M., Lee, L.S. and Henry, R.J. Analysis of SSRs derived from grape ESTs. **Theor. Appl. Genet.** 100 (2000) 723-726.

16. Eujayl, I., Sorrells, M., Baum, M., Wolters, P. and Powell, W. Assesment of genotypic variation among cultivated durum wheat based on EST-SSRs and genomic SSRs. **Euphytica** 119 (2001) 39-43.
17. Kota, R., Varshney, R.K., Thiel, T., Dehmer, K.J. and Graner A. Generation and comparison of EST-derived SSR and SNP markers in barley (*Hordeum vulgare* L.). **Hereditas** 135 (2001) 141-151.
18. Cordeiro, G., Casu, R., McIntyre, C., Manners, J. . and Henry, R.J. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross-transferable to erianthus and sorghum. **Plant Sci.** 160 (2001) 1115-1123.
19. Eujayl, I., Sorrells, M., Baum, M., Wolters, P., and Powell, W. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. **Theor. Appl. Genet.** 104 (2002) 399-407.
20. Miller, R.T., Christoffels, A.G., Gopalakrishnan, C., Burke, J., Ptitsyn, A. A., Broveak, T.R. and Hide, W.A. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. **Genome Res.** 9 (1999) 101-113.
21. Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D. and Waugh, R. Computational and experimental characterization of physically clustered simple sequence repeats in plants. **Genetics** 156 (2000) 847-854.
22. Akagi, H., Yokozeki, Y., Inagaki, A. and Fujimura, T. Microsatellite DNA markers for rice chromosomes. **Theor. Appl. Genet.** 93 (1996) 1071-1077.
23. Metzgar, D., Bytof, J., and Wills, C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. **Genome Res.** 10 (2000) 72-80.
24. Morgante, M. and Olivieri, A.M. PCR-amplified microsatellites as markers in in plant genetics. **Plant J.** 3 (1993) 175-182.
25. Powell, W., Machray, G.C. and Provan, J. Polymorphism revealed by simple sequence repeats. **Trends Plant Sci.** 1(1996) 215-222.
26. Chin, E.C.L. Maize simple repetitive DNA sequences: abundance and allele variation. **Genome** 39 (1996) 866-873.
27. Morgante, M., Hanafey, M. and Powell, W. Microsatellites are preferentially with nonrepetitive DNA in plant genomes. **Nature Genet.** 30 (2002) 194-200.
28. Kantety, R.V., Rota, M.L., Matthews, D.E. and Sorrells, M.E. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. **Plant Mol. Biol.** 48 (2002) 501-510.

Tab.1. Abundance of microsatellites in ESTs of different cereal species

	Barley (IPK)	Barley (Public)	Oats	Maize	Rice	Rye	Sorghum	Wheat
Total number of sequences examined	41600	87205	488	119158	101727	8034	87216	75247
Total length of examined sequences (bp)	21,336,720	55,084,874	144,258	54,691,836	43,928,218	3,719,623	41,570,568	37,528,718
Number of ESTs containing SSRs	3285	20135	7	36152	54453	660	8607	9825
Number of identified SSRs (excluding monomers)	3154(7.6%)	7001(8.0%)	2(0.4%)	7307(6.1%)	11086(10.9%)	672(8.4%)	7619(8.7%)	6010(8%)
Number of primer pairs for redundant SSRs	2083(70.3%)	4521(64.6%)	1(50%)	4969(68.0%)	6855(61.8%)	362(53.9%)	5381(70.6%)	3475(57.8%)
Number of primer pairs for non-redundant SSRs ¹	914(43.8%)	1839(40.7%)	1(100%)	1453(29.2%)	2589(37.8%)	246(67.9%)	1779(33.1%)	1343(38.6%)
Unique primer pairs for total EST sequences (%)	2.2	2.1	0.2	1.2	2.5	3.1	2.1	1.8

¹Primer pairs for development of unique markers for mapping non-redundant genes (unigenes). Primer sequences are available from the corresponding author on request.

Tab. 2. Distribution of SSRs into different classes in different cereal species

Repeat length	Barley			Maize			Rice			Rye			Sorghum			Wheat		
	NN	NNN	NNNN	NN	NNN	NNNN	NN	NNN	NNNN	NN	NNN	NNNN	NN	NNN	NNNN	NN	NNN	NNNN
<10 unit	81.97	99.56	98.85	89.27	99.57	100	83.4	99.68	99.16	87.93	99.42	100	83.36	99.33	96.49	74.53	95.11	98.11
>10 unit	18.03	0.44	1.15	10.73	0.43	0	16.6	0.32	0.84	12.07	0.58	0	16.64	0.67	3.51	25.47	4.89	1.89