American Journal of Plant Sciences, 2015, 6, 3244-3250
Published Online December 2015 in SciRes. http://www.scirp.org/journal/ajps
http://dx.doi.org/10.4236/ajps.2015.619315

In Silico Exploration of *Cannabis sativa* L. Genome for Simple Sequence Repeats (SSRs)

Incoronata Galasso, Elena Ponzoni

Istituto di Biologia e Biotecnologia Agraria (IBBA-CNR), Milan, Italy
Email: galasso@ibba.cnr.it

Received 19 November 2015; accepted 15 December 2015; published 18 December 2015

Abstract

Simple sequence repeat (SSR) or microsatellite markers, are a valuable tool for several purposes such as evaluation of genetic diversity, fingerprinting, marker assisted selection, and breeding. Recent developments in sequencing technologies and bioinformatics analyses provide new opportunity to produce a high number of less costly SSRs. Here, we used for the first time a whole-genome shotgun sequencing of the nuclear genome and transcriptome of hemp to develop microsatellite markers for *C. sativa* L. (hemp). Hemp is an ancient crop that is widely cultivated as a source of fiber, seeds and medicine. The analysis using the MISA program revealed a total of 407,491 SSRs (from mono-nucleotide to deca-nucleotide) in the hemp genome and 15,655 SSRs in the transcriptome. Analysis of the frequency and distribution of SSRs showed that the mono-nucleotide repeats were the most abundant (55.4%) in the genome whereas the tri-nucleotide motifs (30.4%) resulted highly predominant in the transcriptome. Poly A/T was predominant over poly G/C in both genome and transcriptome sequences. Among the tri-nucleotide repeats AAG/CTT (34.5%) resulted the most abundant in the transcriptome. Repeats larger than tri-nucleotide were also observed in the hemp genome and transcriptome. Dinucleotide and tri-nucleotide repeat expansion of 8605 and 1401 times iteration were observed however, other SSR expansion more than 387 times repetition was not found. Primers were designed for amplification of few long microsatellite sequences which could be used to identify polymorphism and to study genetic diversity among hemp cultivars.

Keywords

Microsatellite, Relative Density, Relative Abundance, PCR Amplification

1. Introduction

Repetitive elements are present in large quantities in eukaryotic genome, both in coding and non-coding region

How to cite this paper: Galasso, I. and Ponzoni, E. (2015) *In Silico* Exploration of *Cannabis sativa* L. Genome for Simple Sequence Repeats (SSRs). *American Journal of Plant Sciences*, **6**, 3244-3250. http://dx.doi.org/10.4236/ajps.2015.619315

[1]. Among them the tandemly repeated DNA sequences of 1 - 6 bp are referred to as simple sequence repeats (SSRs), sequence tagged sites (STS) or microsatellites and resulted very useful for genetic marker development and genome application [2] [3]. Simple sequence repeats are codominant, abundant, multi-allelic, and uniformly distributed over the genome, and can be detected by simple reproducible assays [4]. Traditionally, SSRs have been isolated from partially digested genomic DNA libraries and several thousands of clones were screened through colony/plaque hybridization using repetitive DNA probes. Later on several other methods have been used in order to decrease the time and cost invested and simultaneously increasing the yield of microsatellites. Today the increasing whole-genome sequences of many plant species provide sources for SSR mining *in silico*. Therefore, the low cost of *in silico* mining and high abundance of microsatellites in different sequence resources make this approach extremely attractive for the generation of microsatellite markers.

Recently, a whole-genome shotgun sequencing of the nuclear genome and transcriptome of hemp has been reported by van Bekel *et al.* (2011) [5]. This project provides the assembled draft genome and transcriptome of *Cannabis sativa* strain Purple Kush (PK). The contig assembly contains 534.0 Mb without gaps and 786.6 Mb including gaps representing an estimated 65% and 96% genome coverage of the haploid hemp genome ~820 Mb [5]. A total of 136,290 scaffolds were obtained from the whole-genome shotgun assembly and 40,224 from the transcriptome. Availability of hemp genome led to the possibility of *in silico* analysis of the genome for the identification of microsatellite which could be useful for cultivar identification, mapping and genetic diversity evaluation. Therefore, in the present study, we analysed the hemp genome and transcriptome sequences using several publicly available software programs with the objectives: a) to retrieve and characterize microsatellite loci from the genome and transcriptome, b) to develop and characterize a collection of SSR-markers for hemp in terms of frequency, information content, genomic distribution, and c) to assess their potential for diversity analysis in a reference set of hemp cultivars of different origin.

2. Material and Methods

2.1. Identification of Microsatellites

Genomic and transcriptomic sequences of hemp in FASTA format were downloaded from the Cannabis Genome Browser http://genome.ccbr.utoronto.ca/ database. The Perl script MIcroSAtellite (MISA) (http://pgrc.ipk-gatersleben.de/misa/) was used to identify microsatellites from both genomes and coding DNA sequences (CDS) from the transcriptome. To identify the presence of SSRs, only 1 to 10 nucleotide motifs were considered, and the minimum repeat unit was defined as 10 for mono-, 6 for di-, 5 for tri-, tetra-, penta-, hexa-, 3 for septa- and 2 for octa- to deca-nucleotides. Compound SSRs were defined as ≥2 SSRs interrupted by ≤100 bases [6].

The categorization proposed by Weber (1990) [7] was used. Perfect repeats are formed from identical repetitive units; imperfect repeats are units with small mutations, and repetitive compound elements are composed of sequences in which two or more repetitions (perfects or imperfects) are arranged successively with or without nucleotide bases between them.

2.2. Statistical Analysis

SSR types were analysed for their abundance and density per Mb for both genome and coding sequences. Statistical data not present in the MISA output files, like e.g. the relative abundance and the relative density have been calculated using the custom program statistics_misa.py and statgetlongest.py. The relative abundance and density were calculated by following formulas:

$$\text{Relative abundance} = \text{Number of SSRs} / \text{Length of sequence analysed } (\text{Mb})$$

$$\text{Relative density} = \text{Length of SSR } (\text{bp}) / \text{Length of sequence analysed } (\text{Mb})$$

2.3. Sequence Analysis for Primer Designing

Genomic and CDS SSRs generated by MISA were analysed for designing primers flanking the repeats. Genomic microsatellites have been selected that match the following criteria: minimum and maximum repeat length of 30 and 200 bp, respectively and having an up- and downstream flanking region of at least 200 bp. For CDS mi-

crosatellites the minimum and maximum repeat length was set to 20 and 200 bp, respectively with an up- and downstream flanking region of at least 150 bp.

In order to find microsatellites matching the before mentioned criteria the custom programs filterrepeatsmisa.py and getsequences.py were used. The custom programs used in this study (PySSRstat) have been written in the Python 3 language and are available from http://www.nenno.it/PySSRstat.

2.4. Designing SSR Based Primers and Validation of SSR Markers for Amplification

To design primers flanking the microsatellite loci, Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) program was used. The length of the amplicons was set to 100 - 350 bp. Oligonucleotide parameters for Primer3 were set to a length of 18 - 27 bp with an optimum of 20 bp, a GC content of 20% - 80% with an optimum of 50%, a melting temperature (Tm) of 57˚C - 63˚C with an optimum of 60˚C, and a primer Tm maximum difference of 1˚C or 2˚C.

Ten cultivars of industrial non-drug hempseed, which are the most cultivated in Europe (Eletta Campana, Kc Dora, Codimono, Carmaleonte, Felina, Fibranova, Fedora, Futura, Carmagnola and Finola), were chosen and used for the validation of 15 SSR markers randomly selected. Ten SSR markers were chosen from the genomic DNA and five from the transcriptome (**Table 1**). Genomic DNA from all hemp cultivars was isolated from young leaves. Each PCR reaction was performed in a total volume of 15 µl containing 10 ng of genomic DNA, 5 pmole each of forward and reverse primers, 0.1 mM dNTPs, 1 × PCR buffer (10 mM Tris, pH 8.0, 50 mM KCl and 50 mM ammonium sulphate), 1.8 mM $MgCl_2$, and 0.2 unit of Taq DNA polymerase. The cycling conditions involved initial denaturation at 94˚C for 4 min, followed by 36 cycles of denaturation at 94˚C for 1 min, primer annealing at 56˚C for 45 sec, and primer extension at 72˚C for 45 sec. A final extension at 72˚C for 7 min was done and products stored at 4˚C until electrophoresis. The PCR products were resolved by electrophoresis in 2% agarose gels in 1 × TAE buffer and visualized by ethidium bromide staining.

3. Results and Discussion

The analysis by the MISA program revealed a total of 407,491 SSRs (from mono-nucleotide to deca-nucleotide) in the hemp genome and 15,655 SSRs in the transcriptome (**Table 2**). The relative density and abundance of SSRs for the genome was 1527 bp/Mb and 518 SSR/Mb, respectively and for the CDS 1351 bp/Mb and 385 SSR/Mb, respectively (**Table 2**). The relative abundance of SSR/Mb in the hemp genome is in line with that reported by Sonah *et al.*, 2011 [6] for other dicot plant species such as *Arabidopsis thaliana* (416.6/Mb), *Medicago truncatula* (405.8/Mb) and *Populus trichocarpa* (667.9/Mb).

Using MISA program, we obtained a detailed analysis of the frequency and distribution of all mono- to deca-nucleotides repeats from the hemp genomic DNA and CDS (**Table 3**). Similarly to other plant genomes studied so far [6] also in hemp genome the most frequent microsatellite type was the mono-nucleotide repeat (55.4%), whereas the most abundant repeat in the CDS resulted the tri-nucleotide repeats (30.4%) followed by the mono-dinucleotide repeat (27.3%) (**Table 3**). The accumulation of tri-nucleotide repeats in the hemp CDS is consistent with the results of other authors which analysed the CDS of several plant species [6] [8] [9].

Among the other repeats the octa-nucleotide showed the highest frequency for both CDS and genomic DNA, 11.4% and 12.9%, respectively. Except the nona-nucleotide repeat which was 8.3% and 5.8% in the CDS and genome respectively, all the remaining repeats (tetra-, penta- hexa-, septa- and deca-nucleotide) were present below 2.5% (**Table 3**).

Among the mono-nucleotide repeats, the motif A/T was the most common both in the hemp genome and CDS (**Table 4**). The AT/AT di-nucleotides was the most frequent in the genome with 51.1% whereas AG/CT motif was the most abundant in the CDS with 64.6%. For tri- and tetra-nucleotides, the motifs AAT/ATT reached 45.8%, AAG/CTT 34.5%, AAAT/ATTT 54.5%, and AAAG/CTTT 31.6%, respectively. In the hemp genome the penta- to septa-nucleotide repeats were represented by AATAC/GTATT (23.2%), AATGGG/CCCATT (13.0%), and AAAAAAT/ATTTTTT (15.6%), whereas in the transcriptome 16.7% by the penta-nucleotides AAGAG/CTCTT, AACTC/GAGTT and 14.5% by the deca-nucletide AAAGAGAGAG/CTCTCTCTTT.

All the remaining motifs were less than 10% (**Table 4**). As reported by Grover *et al.*, 2007 [10] also in hemp genome and transcriptome, microsatellites show a decrease in abundance with increasing repeat length. In hemp genome the longest mono-nucleotide repeat was Poly A repeated 294 times followed by Poly T iterated 113 times, similarly in the hemp CDS the longest mono-nucleotide repeat was Poly A repeated 47 times followed by

Table 1. Identification number (N), primer sequence and melting temperature (Tm) of primer designed to PCR amplify hemp SSRs.

Identification N.	Primer Name	Sequence	Tm	Size	Repeat
Genome					
Scaffold5651	Scaf5651F	5'GTGGTGGCATCATTCAACAG3'	59	229	(TGG) 10
	Scaf5651R	5'CAAAGCCAAAACTCCCAAAA3'	60		
Scaffold30053	Scaf30053F	5'TGTTGGGTTAAGGGCATTTT3'	59	239	(TC) 24
	Scaf30053R	5'CCTTGTTCTAGCTGCCTTCG3'	60		
Scaffold12289	Scaf12289F	5'GGTGCATTGCAAGAGAACAA3'	59	181	(GA) 18
	Scaf12289R	5'CCCTCAATCCACTCTGAAAAA3'	59		
Scaffold103666	Scaf103666F	5'AGCTTTCGAATTCGTCTGGA3'	59	198	(GAT) 12
	Scaf103666R	5'TCACTCCCATCATTAACCAACTC3'	60		
Scaffold138656	Scaf138656F	5'TGGTCCACCAGGTCAAGATT3'	60	209	(CAA) 14
	Scaf138656R	5'ATTCCCAACTCCTCCGTTCT3'	59		
Scaffold39138	Scaf39138F	5'CTGTCATCACAACCCACCAT3'	59	210	(TGA) 14
	Scaf39138R	5'ACCGATTTCTCCATTGTTGC3'	59		
Scaffold42744	Scaf42744F	5'TTCATCTAGCTGATCTGGCAAA3'	59	215	(TTG) 11
	Scaf42744R	5'CCAACCTCAACTCTCTTCTTCC3'	59		
Scaffold143652	Scaf143652F	5'TGTTGGCGATATTTCCACAGT3'	60	169	(TTG) 12
	Scaf143652R	5'GGGAAAATCATGTCTGCTCAA3'	60		
Scaffold27728	Scaf27728F	5'GCCAAAAATCAAGCAATTCA3'	58	202	(GA) 20
	Scaf27728R	5'GCCCTTGTTTGAGTTTGGAA3'	60		
Scaffold104423	Scaf104423F	5'TGGCCTAACACACTTGCGTA3'	60	245	(TTCT) 12
	Scaf104423R	5'CACCACTTAGAGTTTTGAGTGCTTT3'	60		
Transcriptome					
PK24944	PK24944F	5'GATCCGACTTCCTGATTTCAA3'	59	238	(AAG) 11
	PK24944R	5'ACGTTTGTGGAAGCAAGAGC3'	60		
PK14152	PK14152F	5'CCTCCGATTTGATGCTCATT3'	60	213	(AG) 25
	PK14152R	5'CAAACACTGGTTCAGCCTCA3'	59		
PK18141	PK18141F	5'GAAGAACACGCCAAATCCTC3'	59	245	(ACC) 11
	PK18141R	5'TGAAACTCATCGTCGTCTCG3'	59		
PK13506	PK13506F	5'ACATTTGTGGATGGGGGTAA3'	59	212	(AG) 17
	PK13506R	5'GAACCAGCTTTGGAAACCAT3'	59		
PK27965	PK27965F	5'CCCACCTCCTTCTCCTCTTC3'	60	237	(CTCCCA) 7
	PK27965R	5'TTGAGGCATGGTATTGGTGA3'	59		

Table 2. Number and distribution of SSRs in whole-genome and transcriptome of hemp.

	Genome	Transcriptome
Total size covered by examined sequences (Mb)	786.6	40.63
Total number of sequences examined	136,290	40,224
Total number of SSR identified	407,491	15,655
Total length of SSR (bp)	1,200,858	54,896
Total relative abundance (SSR/Mb)	518	385
Total relative density (bp/Mb)	1527	1351

Table 3. Distribution of SSR motifs in the whole-genome and transcriptome of hemp.

	Motif length	Number	Frequency %	Longest SSR motifs
Genome	Mono	225,883	55.4	$(A)_{294}$, $(T)_{113}$, $(C)_{23}$, $(G)_{24}$
	Di	60,704	14.9	$(GT)_{8605}$
	Tri	28,125	6.9	$(TTA)_{1401}$
	Tetra	2991	0.7	$(AAGA)_{182}$
	Penta	526	0.1	$(ATCCA)_{99}$
	Hexa	362	0.1	$(ACACAT)_{387}$
	Septa	2634	0.6	$(GAGCAAG)_{106}$
	Octa	52,586	12.9	$^{*}(AAAAAAAC)_4$, $(ACCACCAC)_4$, $(CCTCACTC)_4$
	Nona	23,500	5.8	$(TCTCATCAG)_{39}$
	Deca	10,180	2.5	$(AGTGCTAGGT)_4$, $(CTCTCTCGAA)_4$
Transcriptome	Mono	4281	27.3	$(A)_{47}$, $(T)_{43}$, $(G)_{10}$
	Di	2884	18.4	$(AG)_{25}$
	Tri	4762	30.4	$(AGA)_{16}$, $(ATA)_{16}$, $(ATG)_{16}$, $(CAA)_{16}$, $(TCT)_{16}$
	Tetra	187	1.2	$(AGAA)_7$, $(AGGA)_7$, $(TAGA)_7$, $(TTCT)_7$, $(TTTC)_7$
	Penta	36	0.2	$^{**}(AACTC)_6$, $(AGAAG)_6$, $(CTCAA)_6$
	Hexa	47	0.3	$(GATGGT)_8$
	Septa	114	0.7	$(TCCTTGC)_7$
	Octa	1792	11.4	$(CCTCACTC)_4$, $(TTTCTTTT)_4$
	Nona	1303	8.3	$^{***}(AATGATGAT)_3$, $(ACACCAAGA)_3$, $(CAACCAAAC)_3$
	Deca	249	1.6	$^{****}(AAAAAAAAAC)_2$, $(AAAAAAAAAG)_2$, $(AAAAAAGAAA)_2$

*Three out of twenty-three SSRs; **Three out of nine; ***Three out of ten; ****Three out of one hundred eighty-three.

Table 4. The most abundant repeat types from mono- to deca-nucleotide. Freq = Frequency.

	Genome		Transcriptome	
Type	Repeat	Freq.	Repeat	Freq.
Mono	A/T	99.5%	A/T	100.0%
Di	AT/AT	51.1%	AG/CT	64.6%
Tri	AAT/ATT	45.8%	AAG/CTT	34.5%
Tetra	AAAT/ATTT	54.5%	AAAG/CTTT	31.6%
Penta	AATAC/GTATT	23.2%	AAGAG/CTCTT, AACTC/GAGTT	16.7%
Hexa	AATGGG/CCCATT	13.0%	ACCATC/GATGGT, AAGATG/CATCTT	8.5%
Septa	AAAAAAT/ATTTTTT	15.6%	AAAAAAG/CTTTTTT, AAGAGAG/CTCTCTT	7.9%
Octa	AAAAAAAT/ATTTTTTT	6.6%	AAAAAAAG/CTTTTTTT	6.1%
Nona	AAAAAAAAT/ATTTTTTTT	6.5%	AAGAGAGAG/CTCTCTCTT	2.6%
Deca	AAAAAAAAAT/ATTTTTTTTT	6.4%	AAAGAGAGAG/CTCTCTCTTT	14.5%

Poly T iterated 43. The longest di-nucleotide repeat in hemp genome was made of GT/AC repeated 8605 times (scaffold81868) whereas in the hemp CDS was AG/CT repeated 25 times (PK14152). Tri-nucleotide repeats were the first most abundant SSRs present within the hemp CDS and of the 64 triplet repeat types five: (ATG, PK16635), (ATA, PK09074), (TCT, PK14855), (CAA, PK15453), (AGA, PK13649) were made by 16 repeats, while in the genome the longest TTA tri-nucleotide was repeated 1401 times (scaffold120259) (**Table 3**).

Analysing of the 407,491 (genomic SSR) and 15,655 (CDS SSR) repeat motifs using the custom programs filterrepeatsmisa.py and getsequences.py revealed 3353 (0.82%) and 507 (3.24%) repeat motifs, respectively having an up- and downstream flanking region of at least 200 bp for the genomic SSRs and 150 for the CDS SSRs (http://www.hempssr.altervista.org/). The rationale for screening all SSRs generate by MISA using the above programs was necessary in order to capture individual microsatellites along with enough flanking sequence for the design of forward and reverse primers for PCR amplification. However using less stringent parameters probably the number of SSRs will increase.

Among all sequences reported (http://www.hempssr.altervista.org/), fifteen sequences (from genomic and CDS DNA) were randomly chosen to design primers flanking di-, tri-, tetra-, and hexa-nucleotide repeats (see **Table 1**) and validated by PCR. After PCR amplification all SSRs tested showed a prominent PCR product on the agarose gel (**Figure 1(a)**). Furthermore to analyse the potential of these markers for genetic variability studies four of them were tested on ten hemp cultivars. In **Figures 1(b)-(e)** is reported the PCR products after amplification. Although we tested only 4 SSRs the CDS-SSRs appeared more polymorphic than the genomic-SSRs (**Figure 1(d)** and **Figure 1(e)**).

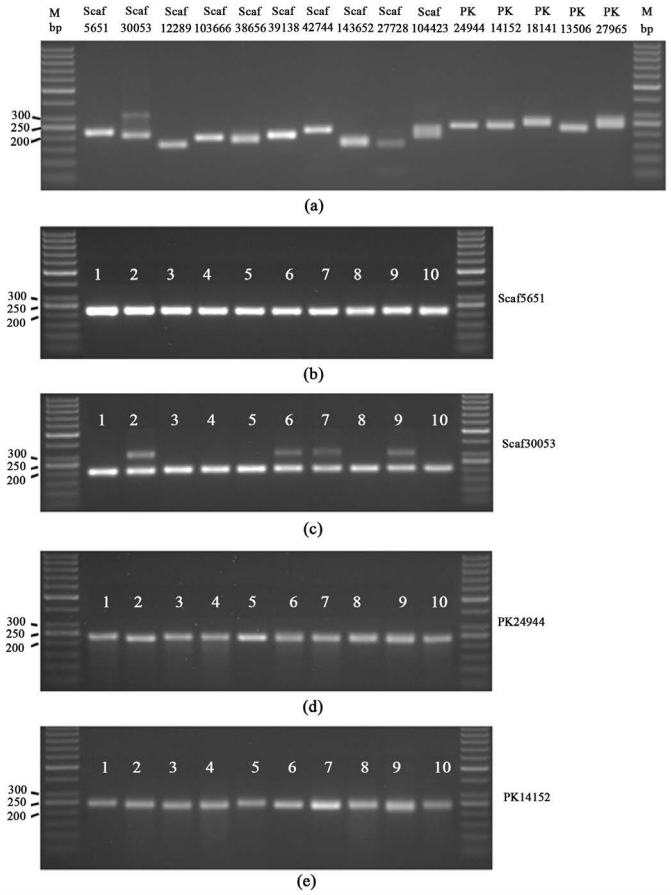

Figure 1. (a) PCR-amplified product of 15 SSR markers tested on the *C. sativa* cv. Futura. PCR-amplified product of Scaf5651 (b); Scaf30053 (c); PK24944 (d) and PK14152 (e) tested on ten hemp cultivars: 1 Eletta Campana; 2 Kc Dora; 3 Codimono; 4 Carmaleonte; 5 Felina; 6 Fibranova; 7 Fedora; 8 Futura; 9 Carmagnola; 10 Finola. M = Molecular marker size in base pair (bp).

4. Conclusion

Traditionally, SSR loci have been isolated from partially digested genomic DNA libraries of small size inserts. Conversely, as showed in this study, the use of whole-genome sequence provided a valuable resource for the development of SSR-markers saving both cost and time, once a sufficient amount of sequences are available. Indeed, in this study, we reported the analysis of whole-hemp genome sequence and the identification of many SSRs that would serve as an important resource for genetic studies.

Acknowledgements

This study was partially supported by Regione Lombardia/CNR. Research Project FilAgro "Strategie innovative e sostenibili per la filiera agroalimentare" Accordo Quadro N.18093/RCC del 5/8/2013. We would like to thank Dr. Mario G. Nenno for writing the custom programs (PySSRstat).

References

[1] Zane, L., Bargelloni, L. and Patarnello T. (2002) Strategies for Microsatellite Isolation: A Review. *Molecular Ecololy*, **11**, 1-16. http://dx.doi.org/10.1046/j.0962-1083.2001.01418.x

[2] Tautz, D. (1989) Hypervariability of Simple Sequences as a General Source for Polymorphic DNA Markers. *Nucleic Acids Research*, **17**, 6463-6471. http://dx.doi.org/10.1093/nar/17.16.6463

[3] Morgante, M. and Olivieri, A.M. (1993) PCR-Amplified Microsatellites as Markers in Plant Genetics. The *Plant Journal*, **3**, 175-182. http://dx.doi.org/10.1111/j.1365-313X.1993.tb00020.x

[4] Powell, W., Machray, G. and Provan, J. (1996) Polymorphism Revealed by Simple Sequence Repeats. *Trends Plant Science*, **1**, 215-222. http://dx.doi.org/10.1016/S1360-1385(96)86898-0

[5] van Bakel, H., Stout, J.M., Cote, A.G., Tallon, C.M., Sharpe, A.G., Hughes, T.R. and Page, J.E. (2011) The Draft Genome and Transcriptome of *Cannabis sativa*. *Genome Biology*, **12**, R102. http://dx.doi.org/10.1186/gb-2011-12-10-r102

[6] Sonah, H., Deshmukh, R.K., Sharma, A., Singh, V.P. and Gupta, D.K. (2011) Genome-Wide Distribution and Organization of Microsatellites in Plants: An Insight into Marker Development in Brachypodium. *PLoS ONE*, **6**, e21298. http://dx.doi.org/10.1371/journal.pone.0021298

[7] Weber, J.L. (1990) Informativeness of Human (dC-dA)n (dG-dT)n Polymorphisms. *Genomics*, **7**, 524-530. http://dx.doi.org/10.1016/0888-7543(90)90195-Z

[8] Morgante, M., Hanafey, M. and Powell, W. (2002) Microsatellites Are Preferentially Associated with Nonrepetitive DNA in Plant Genomes. *Nature Genetics*, **30**, 194-200. http://dx.doi.org/10.1038/ng822

[9] Li, Y.C., Korol, A.B., Fahima, T. and Nevo, E. (2004) Microsatellites within Genes: Structure, Function, and Evolution. *Molecular Biology and Evolution*, **21**, 991-1007. http://dx.doi.org/10.1093/molbev/msh073

[10] Grover, A., Aishwarya, V. and Sharma P.C. (2007) Biased Distribution of Microsatellite Motifs in the Rice Genome. *Molecular Genetics and Genomics*, **277**, 469-480. http://dx.doi.org/10.1007/s00438-006-0204-y