*Available online www.jsaer.com*

***Journal of Scientific and Engineering Research*, 2017, 4(11):154-163**


**Research Article**

**ISSN: 2394-2630**
**CODEN(USA): JSERBR**

# Microsatellite Distribution and Densities in Promoter and Gene Body Entities of Some Plant Genes

**Mehmet Karaca[1], Ayse Gul Ince[2], Emine Uygur Gocer[1], Adnan Aydin[1]**

[1]Field Crops Department, Faculty of Agriculture, Akdeniz University, 07070 Antalya, Turkey
[2]Vocational School of Technical Sciences, Akdeniz University, 07070 Antalya, Turkey

**Abstract** Advances in sequencing and computational technologies evoked tremendous amount of genomic and transcriptomic data. Annotated genomic and transcriptomic DNA sequence data make it much easier for researchers to view, sort and analyze sequence contents. Annotated data containing information for regulatory (enhancers, promoters and introns) and gene body entities (exons and untranslated regions, UTRs) could be effectively used to obtain and characterize the distribution and densities of microsatellites. Objectives of this study were to determine distribution and density of microsatellites or microsatellite motif lengths (mono-, di-, tri-, tetra-, penta- and hexa-nucleotides) between each gene entity (promoter, exon, 5'UTR, intron and 3'UTR), and between a gene entity and total entities. This study used 809,682 gene entities extracted from 203,758 unique sequences obtained from nucleotide databases. Results showed that there were 15,321 microsatellites in gene entities. Distribution and density of most microsatellites located in gene entities were statistically significant. Microsatellite densities of promoters were the highest while exons contained the least amount of microsatellites. UTRs at the 5'- terminals contained the second highest microsatellite densities after promoters. The distribution and density of microsatellite motifs were also statistically different among gene entities. Exons contained the highest tri-nucleotides while promoters contained di-nucleotides. Results clearly revealed that microsatellites display nonrandom distribution between promoter and gene body entities confirming that microsatellites involve in transcriptional and translational regulation in plants. Our study also indicated that microsatellite genetic markers located on transcriptomic data definitely lack valuable allelic variations presented in enhancers, promoters, introns and intergenic regions.

**Keywords** exons, introns, promoters, simple sequence repeats, tandem repeats, UTR

## 1. Introduction

Eukaryotic genomes contain significant amount of tandem and non-tandem repeats. Tandem repeats (TRs) are DNA sequence motifs that contain at least two adjacent repeating units. TRs exist in both prokaryotic and eukaryotic genomes [1, 2]. Three categories are given to distinguish TRs based on different repeat unit size and repeating times: (i) microsatellites, also called simple sequence repeats (SSRs) consist of unit size 1–6 bp, repeating 5-100 times, frequently found in euchromatic regions; (ii) minisatellites (also called variable number of tandem repeats, VNTRs) consist of unit size 10–400 bp, repeating 20-50 times, generally found in euchromatic and heterochromatic regions, and (iii) satellites consist of unit size 5-300 bp, repeating 10,000 - 1,000,000 times, generally found in heterochromatic regions of centromeres and telomeres [1, 3-6].

There still exist scientific debates on tandem repeats regarding to their function and occurrence within genomes. There is no scientific consensus about their distributions (random vs. nonrandom), functionalities and occurrences within regulator regions such as promoters and gene entities (gene bodies) such as exon, intron, 5'- and 3'-untranslated terminal region (UTR). Among TRs, microsatellites have been the most widely used marker

type for genotyping plants over the past 20 years. Microsatellites are highly informative, co-dominant, multi-allelic genetic markers that are experimentally reproducible and transferable among related plant species. Microsatellites are useful for genetic studies of cultivated or closely related wild plant species because they are multi-allelic, highly polymorphic and follow genetic laws of Mendel. Microsatellites have very high mutation rates (as high as $10^{-4}$–$10^{-3}$ per generation), making them more polymorphic and multi-allelic [5, 7-10].

Repeat polymorphisms in microsatellites evolve through three main processes such as Strand-slippage replication, point mutation, and recombination (unequal crossing-over and gene conversion). Strand-slippage replication is a DNA replication error in which the template and nascent strands are looped out due to mismatches causing to repeat expansion. Also unequal crossing-over and gene conversion lead to microsatellite sequence contractions and expansions. However, in order to generate polymorphisms these mutational processes should escape DNA mismatch repair (MMR) systems. Those mutations that have escaped from the corrections of MMR systems would become new alleles. These new alleles could cause a frame-shift, a fluctuation of gene expression, inactivation of gene activity, and/or a change of function, and eventually phenotypic changes. Due to existence of higher gene conversions in microsatellites, they have been implicated in plant recombination hot spots. These hot spots are known to show nucleosome depletion [11-16].

Many genetic studies have used microsatellites as genetic markers and today microsatellites are widely considered marker of choice in plants. Although microsatellites are considered to be randomly distributed within genomes including gene body entities, their presence on those genes under environmental pressures may cause their biased frequency in a population [17]. Numerous lines of evidence have demonstrated that genomic distribution of microsatellite repeats is nonrandom. Results of Zhao et al. [8] revealed that there existed no clear relationship between tandem repeat density and genome size. They also found that tandem repeats display nonrandom distribution within both intragenic and intergenic regions. However, they did not provide data on distribution of microsatellites and repeat length motifs between promoters and gene body, between gene entities such as introns and exons. Qu and Liu [18] reported that microsatellite densities were found to be highest in 5' untranslated terminal region (UTR), followed by 3'UTR, promoter, intronic, intergenic, and protein coding regions in maize. Microsatellites within genes seem to be subjected to stronger selective pressure than other genomic regions. These microsatellites on genes may provide a molecular basis for fast adaptation to environmental changes.

Objectives of this study were to determine distribution and density of microsatellites or microsatellite motifs (mono-, di-, tri-, tetra-, penta- and hexa-nucleotides) between each gene entity, and between a gene entity and total (combined) entities (promoter, exon, 5'UTR, intron and 3'UTR) using 809,682 gene entities extracted from 203,758 unique sequences obtained from NCBI databases.

## 2. Materials and Methods

### 2.1. DNA Sequence Data

A total of 203,758 GenBank formatted sequences consisting of monocotyledons and dicotyledons were obtained from publicly available NCBI data (ftp://ftp.ncbi.nih.gov/) and used in the present study. EpiOne software [19] was used to extract promoter and gene body entities using algorithms under the promoters & gene body options enabled us to collect promoter, 5'UTR, exon, intron and 3'UTR entities.

### 2.2. Microsatellite Analysis

Microsatellites in promoter, 5'UTR, exon, intron and 3'UTR entities were identified using the Tandem Repeats Analyzer 1.5 (TRA1.5) software [20]. Microsatellites (SSRs) in the present study were considered sequences containing a minimum of 18, 9, 7, 5, 5 and 4 nucleotide repeats for mono-, di-, tri-, tetra-, penta- and hexa-nucleotides, respectively. These repeat criteria were chosen since they are commonly used in other plant species [21, 22].

### 2.3. Statistical Analysis

Chi-square ($\chi^2$) goodness-of-fit tests with 1 degree of freedom were applied to test whether microsatellite densities were significantly different within and between promoter and gene body entities and among gene body entities. Following formula was used:

Where; $E_i$ is the expected number of microsatellites in a dataset; N is the total number of

$E_i = \frac{N}{L} \times L_i$ microsatellites in the two different datasets; L is the total length in base pairs of the two datasets; and Li is the length in base pairs of the dataset under investigation [22].

## 3. Results and Discussion

In the present study, 530,868 DNA sequences from a large number of plant species were used. EpiOne software [19] was utilized to extract promoters and gen body entities from annotated GenBank sequence data. A total of 809,962 entities consisting of promoter, 5'UTR, exon, intron and 3'UTR were identified (Table 1). Among the entities exons were the largest dataset while the 3'UTRs were the least dataset. Each entity sequence of promoter contained one promoter while entities of exon and intron sequences contained more than one entity. For instance, there were more than one exon per sequence with an average 7.2 exons per sequence. More than one exon and intron are expected from a gene sequence, however, we found that there were more than one 5'UTR and 3'UTR in some gene sequences. In the present study, UTR entities occurred more than one per gene sequence were considered different UTRs since they contained their own DNA sequences, although some of which contained overlapped sequences. Total lengths of DNA sequences were539,191,650 nucleotides. TRA 1.5 program [20] identified a total of 15,321 microsatellites and they were presented in Table 1. Table 2 showed number of motifs and dominant motif contents. Table 3 presented density differences between a gene entity and the total entities while Table 4 presented density differences between each gene entity.

### 3.1. Microsatellite Distribution and Density in Promoters

Among 6,335 promoters, 1,069 contained microsatellites indicating that a significant amount of (16.87%) promoters contained microsatellites (Table 1). This also indicated that each 11.2 kilo-bases of promoter sequences contained one microsatellite. This was the highest density of microsatellites among gene entities used in this study. Mono-, di-, tri-, tetra-, penta- and hexa-nucleotides of promoters were different in amounts and repeat contents (Figure 1, Table 2, 3 and 4).

Di-nucleotide repeat motifs were the highest among microsatellites (35.55%) while tetra-nucleotide repeat motifs were the least (6.83%) in promoters. The second most abundant microsatellites were tri- and mono-nucleotide motifs followed by hexa-nucleotide motifs (Table 2). Dominant base compositions of mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide motifs of promoters were T/A (49.2%), AT/AT (26.6%), CCA/TGG (28.2%), AATT/AATT (73.9%), CCGGC/GCCGG (32.51%) and AAAAAG/CTTTTT (56.4%). In promoter sequences, A/T nucleotides were dominant type and were present in each motif from mono- to hexa-nucleotides. Previous studies of Victoria et al. [2] and Zhou et al. [23] revealed that microsatellite motifs of a genome was dependent upon the genome DNA content. For instance, they found GC repeat motifs in genomes with high GC rich genomes. Because plant promoters are rich in A/T contents, it was not surprising to identify promoters rich in A/T content.

Extremely significant ($P \leq 0.0001$) total microsatellite ($\sum$) density difference was identified between promoter and total gene entities (Table 3). We also noted that mono-, di-, tri-, tetra-, penta- and hexa-nucleotides between promoter and total gene entities were extremely significant ($P \leq 0.0001$). Microsatellite densities between promoter and each gene entity were also found extremely significant ($P \leq 0.0001$) as shown in Table 4. Density of mono-nucleotides between promoter and intron was not statistically significant while density of mono-nucleotides between promoter and 5'UTR was statistically significant ($P \leq 0.01$). Densities of mono-nucleotides between promoter and 3'UTR, between promoter and exon were extremely significant ($P \leq 0.0001$) as shown in Table 4. With the exception of tri-nucleotides and tetra-nucleotides between promoter and 3'UTR, and density of tri-nucleotides between promoter and exon, densities of remaining were statistically significant (Table 4).

**Table 1**: Number of GenBank sequences, gene entities and microsatellites used in the present study

| Type of Entity | Number of GI | Number of entity | Length (bp) | Microsatellites # | Microsatellites (%) |
|---|---|---|---|---|---|
| Promoter | 6,335 | 6,335 | 12,800,530 | 1,069 | 16.87 |
| 5'UTR | 2,421 | 6,159 | 2,702,144 | 630 | 10.23 |
| Exon | 69,819 | 502,760 | 260,200,758 | 5,860 | 1.17 |
| Intron | 122,952 | 288,842 | 261,071,604 | 7,522 | 2.60 |
| 3'UTR | 2,207 | 5,586 | 2,416,614 | 240 | 4.30 |
| Total | 203,758 | 809,682 | 539,191,650 | 15,321 | 1.89 |

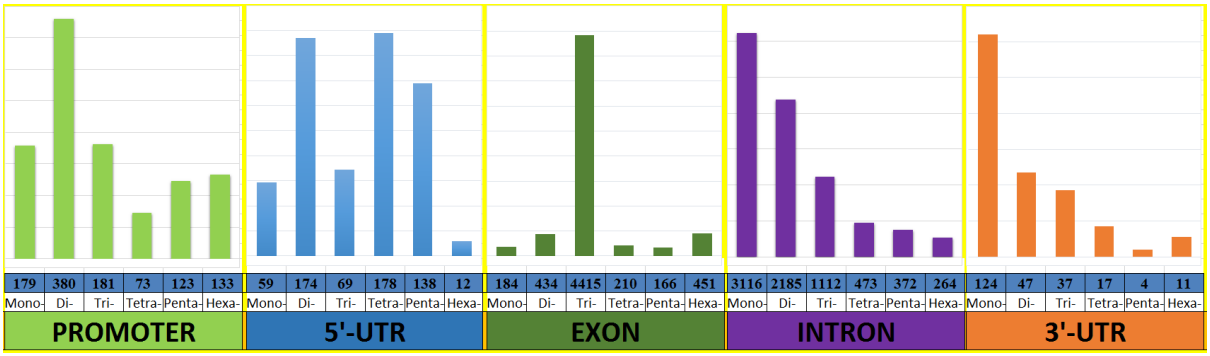


*Figure 1: Graphical illustration of microsatellite distribution within gene entities*

**Table 2**: Microsatellite motif contents in promoters and gene body entities

| Entity | Microsatellite Motif | | | | | |
|---|---|---|---|---|---|---|
| | Mono- | Di- | Tri- | Tetra- | Penta- | Hexa- |
| **Promoter** | 179<br>T/A<br>(49.2%) | 380<br>AT/AG<br>(26.6%) | 181<br>CCA/TGG<br>(28.2%) | 73<br>AATT/AATT<br>(73.9%) | 123<br>CCGGC/GCCGG<br>(32.5%) | 133<br>AAAAAG/CTTTTT<br>(56.4%) |
| **5'UTR** | 59<br>T/A<br>(66.1%) | 174<br>CT/AG<br>(52.8%) | 69<br>AAG/CTT<br>(20.3%) | 178<br>ACAT/ATGT<br>(92.1%) | 138<br>CCCTC/GAGGG<br>(84.1%) | 12<br>AAGCAC/GTGCTT<br>(33.3%) |
| **3'UTR** | 124<br>A/T<br>(88.7%) | 47<br>TA/AT<br>(61.7%) | 37<br>AGG/CCT<br>(32.4%) | 17<br>TATG/CATA<br>(35.3%) | 4<br>ATATG/CATAT<br>(75%) | 11<br>GAGATG/CATCTC<br>(36.4%) |
| **Exon** | 184<br>A/T<br>(61.4%) | 434<br>TC/GA<br>(50.2%) | 4415<br>CGC/GCG<br>(19.5%) | 210<br>ACAT/ATGT<br>(78.1%) | 166<br>CCCTC/GAGGG<br>(69.9%) | 451<br>ACAGCA/TGCTGT<br>(3.10%) |
| **Intron** | 3116<br>A/T<br>(62.8%) | 2185<br>AT/TA<br>(29.4%) | 1112<br>AGA/TCA<br>(17.8%) | 473<br>ATTT/AAAT<br>(23.9%) | 372<br>TAACC/GGTTA<br>(15.3%) | 264<br>TTTTGA/TCAAAA<br>(10.6%) |

**Table 3**: Microsatellite and motif density difference between an entity and the total entities

| Gene entity | Microsatellites | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Bases** | **Mono-** | **Di-** | **Tri-** | **Tetra-** | **Penta-** | **Hexa-** | **∑** |
| Promoter | 12800530 | 179 | 380 | 181 | 73 | 123 | 133 | 1069 |
| Total entities | 526391120 | 3483 | 2840 | 5633 | 878 | 680 | 738 | 14252 |
| Total | 539191650 | 3662 | 3220 | 5814 | 951 | 803 | 871 | 15321 |
| $\chi^2$ | | **99.86** ** | **1234.73** ** | **13.71** ** | **115.35** ** | **580.46** ** | **624.98** ** | **1400.82** ** |
| 5'UTR | 2702144 | 59 | 174 | 69 | 178 | 138 | 12 | 630 |
| Total entities | 536489506 | 3603 | 3046 | 5745 | 773 | 665 | 859 | 14691 |
| Total | 539191650 | 3662 | 3220 | 5814 | 951 | 803 | 871 | 15321 |
| $\chi^2$ | | **90.48** ** | **1552.11** ** | **54.81** ** | **6328.53** ** | **4482.84** ** | **13.42** ** | **4006.12** ** |
| 3'UTR | 2416614 | 124 | 47 | 37 | 17 | 4 | 11 | 240 |
| Total entities | 536775036 | 3538 | 3173 | 5777 | 934 | 799 | 860 | 15081 |
| Total | 539191650 | 3662 | 3220 | 5814 | 951 | 803 | 871 | 15321 |
| $\chi^2$ | | **708.42** ** | **73.83** ** | **4.62** | **38.24** ** | **0.04** | **12.96** ** | **429.42** ** |
| Exon | 260200758 | 184 | 434 | 4415 | 210 | 166 | 451 | 5860 |
| Total entities | 278990892 | 3478 | 2786 | 1399 | 741 | 637 | 420 | 9461 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Total | 539191650 | 3662 | 3220 | 5814 | 951 | 803 | 871 | 15321 |
| $\chi^2$ | | **2741.17** ** | **1559.86** ** | **1783.98** ** | **260.95** ** | **244.71** ** | **4.33** | **614.74** ** |
| Intron | 261071604 | 3116 | 2185 | 1112 | 473 | 372 | 264 | 7522 |
| Total entities | 278120046 | 546 | 1035 | 4702 | 478 | 431 | 607 | 7799 |
| Total | 539191650 | 3662 | 3220 | 5814 | 951 | 803 | 871 | 15321 |
| $\chi^2$ | | **1971.78** ** | **487.14** ** | **1997.52** ** | **0.66** | **1.41** | **114.37** ** | **2.81** |

*: significant P≤0.01, **: extremely significant P≤ 0.0001

Densities of motifs between promoter and other gene entities were statistically significant. The none-significant tri-nucleotide repeats differences between promoter and 3'UTRs,and between promoter and exon could reduce the occurrence for heterochromatin-mediated-like gene silencing and eventually reducing the phenotypic changes [6, 24-27]. Previous studies clearly indicated or revealed that microsatellites in promoter sequences could form unusual secondary structures like H-DNA, G-quadruplex (G4), Z-DNA, and stress-induced duplex destabilized DNA (SIDD) DNA that help or direct transcription control [28, 29]. Microsatellites may influence the chromatin remodeling and accessibility by transcription factors [30, 31]. Also microsatellites appear to be the important components of insulators, silencers and enhancers [12, 29, 32]. Microsatellites located within or vicinity of promoters are considerably more polymorphic than other regions [18, 33, 34]. Our results indicated that genetic markers located in microsatellites of promoters could be developed from promoter sequences since they contain higher occurrences of microsatellites.

**3.2. Microsatellite Distribution and Density in UTRs**

Among 6,159, 630 5'UTRs and among 5,586,240 3'UTRs contained microsatellites indicating that 10.23% and 4.30% 5'UTRs and 3'UTRs contained microsatellites, respectively (Table 1). Analyses indicated that each 4.29 kilo-bases of 5'UTR and each 10.7 kilo-bases of 3'UTR sequences contained one microsatellite. This indicated that 5'UTRs contained about 3 times higher densities of microsatellites than 3'UTR sequences. Motif densities of 5'UTRs and 3'UTRs were different in amounts and contents. Tetra- and di-nucleotide repeats were the highest (28.25% and 27.62%, respectively) motifs while hexa-nucleotide repeats were the least (1.9%) in 5'UTRs. The most abundant motifs were mono-nucleotides (51.67%), followed by di-nucleotides while penta-nucleotides were the least repeats (1.88%) in 3'UTRs (Figure 1, Table 2, 3 and 4).

Dominant base compositions of mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides of 5'UTRs were T/A (66.1%), CT/AG (52.8%), AAG/CTT (20.3%), ACAT/ATGT (92.1%), CCCTC/GAGGG (84.1%) and AAGCAC/GTGCTT (33.3%). Dominant repeat motif contents of 3'UTRs consisted of T/A (88.7%), TA/AT (61.7%), AGG/CCT (32.4%), TATG/CATA (35.3%), ATATG/CATAT (75%) and GAGATG/CATCTC (36.4%) for mono-, di-, tri-, tetra- penta- and hexa-nucleotides, respectively (Table 2).

We compared microsatellite densities between 5'UTR and 3'UTR, between 5'UTR and total gene entities, and between 3'UTR and total gene entities. Total microsatellite (∑) density between 5'UTR and total gene entities, and between 3'UTR and total gene entities were extremely significant (Table 3). Mono-, di-, penta- and hexa-nucleotides of 5'UTRs and 3'UTRs were extremely significant (P≤ 0.0001). On the other hand, penta-nucleotide repeats of 3'UTRs and total gene entities were not statistically significant (Table 3).

Total microsatellite (∑) densities between 5'UTR and 3'UTR, between 5'UTR and intron, between 5'UTR and exon, between 3'UTR and intron, between 3'UTR and exon were extremely significant (P≤ 0.0001, Table 4). Identified significant microsatellite density differences between UTR and other gene entities indicated that microsatellites play some important biological roles in plant gene regulations. Tri-nucleotides and penta-nucleotide densities between 5'UTR and 3'UTR, tri-nucleotide density between 3'UTR and exon, tetra-nucleotide density between 3'UTR and intron were not statistically different (Table 4). Mono-, di-, tetra- and hexa-nucleotide densities between 5'UTR and 3'UTR and between 5'UTR and other gene entities, between 3'UTR and other gene entities were significantly different (P≤ 0.0001) among other gene entities (Table 4).

**Table 4**: Total microsatellite and motif density differences between gene entities

| Gene entity | Bases | Microsatellites | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Mono-** | **Di-** | **Tri-** | **Tetra-** | **Penta-** | **Hexa** | ∑ |
| Promoter | 12800530 | 179 | 380 | 181 | 73 | 123 | 133 | 1069 |
| 5'UTR | 2702144 | 59 | 174 | 69 | 178 | 138 | 12 | 630 |
| Total | 15502674 | 238 | 554 | 250 | 251 | 261 | 145 | 1699 |
| $\chi^2$ | | **8.96** * | **75.21** ** | **17.97** ** | **498.92** ** | **227.82** ** | **8.44** * | **455.84** ** |
| Promoter | 12800530 | 179 | 380 | 181 | 73 | 123 | 133 | 1069 |
| 3'UTR | 2416614 | 124 | 47 | 37 | 17 | 4 | 11 | 240 |
| Total | 15217144 | 303 | 427 | 218 | 90 | 127 | 144 | 1309 |
| $\chi^2$ | | **142.25** ** | **7.59** * | **0.19** | **0.61** | **15.41** ** | **7.32** * | **5.90** |
| Promoter | 12800530 | 179 | 380 | 181 | 73 | 123 | 133 | 1069 |
| Exon | 260200758 | 184 | 434 | 4415 | 210 | 166 | 451 | 5860 |
| Total | 273001288 | 363 | 814 | 4596 | 283 | 289 | 584 | 6929 |
| $\chi^2$ | | **1617.36** ** | **3212.16** ** | **5.79** | **282.10** ** | **927.52** ** | **427.42** ** | **1788.13** ** |
| Promoter | 12800530 | 179 | 380 | 181 | 73 | 123 | 133 | 1069 |
| Intron | 261071604 | 3116 | 2185 | 1112 | 473 | 372 | 264 | 7522 |
| Total | 273872134 | 3295 | 2565 | 1293 | 546 | 495 | 397 | 8591 |
| $\chi^2$ | | **4.26** | **592.04** ** | **252.33** ** | **92.67** ** | **452.19** ** | **740.47** ** | **1163.91** ** |
| 5'UTR | 2702144 | 59 | 174 | 69 | 178 | 138 | 12 | 630 |
| 3'UTR | 2416614 | 124 | 47 | 37 | 17 | 4 | 11 | 240 |
| Total | 5118758 | 183 | 221 | 106 | 195 | 142 | 23 | 870 |
| $\chi^2$ | | **31.00** ** | **59.69** ** | **6.44** | **115.93** ** | **112.29** ** | **0.00** | **134.44** ** |
| 5'UTR | 2702144 | 59 | 174 | 69 | 178 | 138 | 12 | 630 |
| Exon | 260200758 | 184 | 434 | 4415 | 210 | 166 | 451 | 5860 |
| Total | 262902902 | 243 | 608 | 4484 | 388 | 304 | 463 | 6490 |
| $\chi^2$ | | **1291.52** ** | **4549.88** ** | **11.51** ** | **7671.86** ** | **5882.55** ** | **11.13** ** | **4806.19** ** |
| 5'UTR | 2702144 | 59 | 174 | 69 | 178 | 138 | 12 | 630 |
| Intron | 261071604 | 3116 | 2185 | 1112 | 473 | 372 | 264 | 7522 |
| Total | 263773748 | 3175 | 2359 | 1181 | 651 | 510 | 276 | 8152 |
| $\chi^2$ | | **21.77** ** | **938.62** ** | **270.39** ** | **4447.19** ** | **3409.26** ** | **30.07** ** | **3613.22** ** |
| 3'UTR | 2416614 | 124 | 47 | 37 | 17 | 4 | 11 | 240 |
| Exon | 260200758 | 184 | 434 | 4415 | 210 | 166 | 451 | 5860 |
| Total | 262617372 | 308 | 481 | 4452 | 227 | 170 | 462 | 6100 |
| $\chi^2$ | | **5228.06** ** | **413.31** ** | **0.39** | **107.43** ** | **3.83** | **10.81** * | **607.87** ** |
| 3'UTR | 2416614 | 124 | 47 | 37 | 17 | 4 | 11 | 240 |
| Intron | 261071604 | 3116 | 2185 | 1112 | 473 | 372 | 264 | 7522 |
| Total | 263488218 | 3240 | 2232 | 1149 | 490 | 376 | 275 | 7762 |
| $\chi^2$ | | **301.92** ** | **34.70** ** | **67.06** ** | **35.12** ** | **0.09** | **28.76** ** | **404.00** ** |
| Exon | 260200758 | 184 | 434 | 4415 | 210 | 166 | 451 | 5860 |
| Intron | 261071604 | 3116 | 2185 | 1112 | 473 | 372 | 264 | 7522 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Total | 521272362 | 3300 | 2619 | 5527 | 683 | 538 | 715 | 13382 |
| $\chi^2$ | | **2595.26 **** | **1164.84 **** | **1984.97 **** | **100.40 **** | **78.19 **** | **49.53 **** | **200.90 **** |

*: significant P≤0.01, **: extremely significant P≤ 0.0001

In the present study we detected more microsatellites in 5'UTR in comparison to 3'UTR. This indicated that variations of microsatellites in 5'UTRs might have critical biological roles. Microsatellites in 5'UTRs could serve as protein binding sites, thereby regulating gene translation and protein component and function. Previous research also indicated that microsatellite variations in 5'UTRs regulate gene expression by affecting transcription and translation [24, 25, 34-36].

Among the gene body entities, microsatellite densities in 3'UTR sequences ranked the third, indicating the important roles of microsatellites in 3'UTRs. We noted that CAG/CTG repeats were not abundant in 3'UTR sequences. It is known that CAG/CTG expansions and contractions cause RNA slippage during the transcription and lead to transcription of mRNA several kilo-bases longer than the expected size. It is generally assumed that during transcription, transient pausing of the RNA polymerase complex promotes backward slippage and leads to resynthesize of the same RNA sequence. Expansions in the 3'UTRs might cause transcription slippage and produce expanded mRNA, which can disrupt RNA splicing and may disrupt other cellular functions [6, 8, 12].

**3.3. Microsatellite Distribution and Density in Introns**

Among 288,842 introns, 7,522 (2.6%) contained microsatellites (Table 1). This indicated that each 34.71 kilo-bases of intron sequences contained one microsatellite. Intron contained mainly mono-nucleotides (41.43%) followed with di-nucleotides (29%) and tri-nucleotides (14.78%). Hexa-nucleotide repeats were the least (3.51%), followed with tetra-nucleotides (4.95%) and penta-nucleotides (6.29%). As it can be seen in Figure 1 and Table 2 amount of microsatellite decreased as the motif length increased from mono- to hexa-nucleotides.

Dominant mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides in introns were T/A (62.8%), AT/AT (29.4%), AGA/TCA (17.8%), ATTT/AAAT (23.9%), TAACC/GGTTA (15.3%) and TTTTGA/TCAAAA (10.6%) (Table 2). Total microsatellite (∑) density between introns and the total gene entities was not statistically different (Table 3). Also densities of tetra-, and penta-nucleotides between intron and total gene entities were not statistically different. On the other hand, densities of mono-, di-, tri- and hexa-nucleotides between introns and other gene entities were extremely significant (P≤ 0.0001). As shown in Table 2 microsatellite density between intron and exon was extremely significant (P≤ 0.0001). Also densities of mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides between introns and exons were extremely significant (Table 4).

Intronic microsatellites could play a role in the transport and alternative splicing, abnormal splicing and stability and mRNA half-life and in gene silencing. An intronic microsatellites can also behave as a co-regulator with microsatellites in the 5'UTR for gene expression. Intronic polymorphism can result in abnormal splicing. Intronic splicing enhancers have been identified that can mediate tissue-specific exon inclusion [25, 35-37].

**3.4. Microsatellite Distribution and Density in Exons**

Among 502,760 exons, 5,860 (1.17%) contained microsatellites (Table 1). This indicated that each 44.4 kilo-bases of exon sequences contained one microsatellite. This indicated that among the gene entities studied exons contained the least amount of microsatellites. Exons contained mainly tri-nucleotides (75.35%), followed with hexa-nucleotides (7.70%) and di-nucleotides (7.41%). Penta-nucleotide repeats were the least (2.83%), followed with tetra-nucleotides (3.58%) and mono-nucleotides (3.14%). As it can be seen in Figure 1 and Table 2 among exon sequences, the dominant repeat unit sizes were three-fold nucleotides (tri-nucleotides and hexa-nucleotides) because it is assumed that such motifs are selected to avoid frame shift mutations that would affect translation [4, 6, 42].

Dominant mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide motif contents were T/A (61.4%), TC/GA (50.2%), CGC/GCG (19.5%), ACAT/ATGT (78.1%), CCCTC/GAGGG (69.9%) and ACAGCA/TGCTGT (3.1%) in exons (Table 2). Among the repeats in exons, C/G nucleotide repeats were dominant in each motif with exception of mono-nucleotides which contained A/T motifs.

Total microsatellite (∑) density between exons and total gene entities was statistically significant. We also noted that densities of mono-, di-, tri-, tetra- and hexa-nucleotides between exon and total gene entities were also

statistically significant while density of penta-nucleotides between exon and total gene entities were not statistically significant (Table 3). Highly abundant tandem repeats in introns may involve with both constitutive and alternative splicing activities [6, 37-41].

In the present study, the low abundance of microsatellite motifs other than tri-nucleotides in the exonic sequences indicated that those microsatellites are selected against possible frame shift mutations. Previous studies revealed that in *Arabidopsis thaliana* a dramatically expanded TTC/GAA repeats in the intron of the gene encoding the large subunit 1 of the isopropyl malate isomerase cause an environment dependent reduction in the enzyme's activity and severely impairs plant growth. Contraction of the expanded TTC/GAA repeats can reverse the detrimental effect on the phenotype. Interestingly, there are substantial data indicating that microsatellite expansions or contractions in protein-coding regions can lead to a gain or loss of gene function via frame shift mutation or expanded toxic mRNAs [6, 24, 25, 38, 40, 43].

Tri- and hexa-nucleotide repeats in exons appeared to be controlled by stronger mutation pressure than in other gene regions. Microsatellite expansions and contractions in exons are avoided to keep stable protein products. Such a feature can help explain why three-fold nucleotide motifs such tri-nucleotides and hexa-nucleotides are more frequent than others to reduce potential translational frame shifting [24, 25, 40, 42].

Triplet repeats showed approximately two-fold greater frequency in exonic regions than other gene entities. Interestingly AAT motifs in exonic sequences were very low probably due to the fact that TAA-based variants code for stop codons that have a direct effect on protein synthesis in eukaryotes. Microsatellite variations in exons could cause protein functional changes, loss of function, and protein truncation [4, 6, 34, 40].

## 4. Conclusions

Present study confirmed that microsatellites are important components of promoters and gene body entities. There existed significant microsatellite density differences within and between gen entities analyzed in the present study. Promoters, UTRs, and introns, which are three regulatory gene entities, contained higher densities of microsatellites. We noted that exonic sequences contained the highest amount of tri-nucleotide repeats while microsatellite abundance was the least in exons. Results clearly showed that promoters contained the highest amount of microsatellites and microsatellite motifs dominantly consisted of di-nucleotides. Within UTRs, 5'UTR sequences contained the second highest microsatellites after promoters.

Intronic sequences and 3'UTR sequences contained the highest amount of mono-nucleotides. Overall results suggested that expansions or contractions of microsatellites within regulatory regions such as promoters and introns have effects on protein binding including transcription factors, conformation of DNA, nucleosome assemblies, export to cytoplasm, RNA splicing, stability and half-life and tissue specific gene expression. Additionally this study revealed that the biological function of a microsatellite is definitely related to its position in gene entities. Our findings suggested that microsatellite genetic markers could be developed from promoter and 5'UTR sequences since they contain statistically significant microsatellite densities from other gene entities.

## References

[1]. Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Hypervariable Minisatellite Regions in Human DNA. *Nature*, 314:67–73.

[2]. Victoria, F. C., Maia, L. C., & Oliveira, A. C. (2011). In silico Comparative Analysis of SSR Markers in Plants. *BMC Plant Biology*, 11:15.

[3]. Ince, A. G., Karaca, M., & Onus, A. N. (2011). Exact Microsatellite Density Differences Among *Capsicum* Tissues and Development Stages. *Journal of Agricultural Sciences*, 17:291-299.

[4]. Gemayel, R., Boeynaems, J. C. S., &Verstrepen, K. J. (2012) Beyond Junk-Variable Tandem Repeats as Facilitators of Rapid Evolution of Regulatory and Coding Sequences. *Genes*, 3:461–480.

[5]. Ince, A. G., & Karaca, M. (2016). Analysis of Housekeeping and Tissue Specific ESTs for Inexact Microsatellites in *Capsicum* L. *Journal of Scientific and Engineering Research*, 3:663-668.

[6]. Bagshaw, A. T. M. (2017). Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes. *Genome Biology and Evolution*, 9:2428–2443.

[7]. Ananda, G., Walsh, E., Jacob, K. D., Krasilnikova, M., Eckert, K. A., Chiaromonte, F., & Makavo, K. D. (2013). Distinct Mutational Behaviors Differentiate Short Tandem Repeats from Microsatellites in the Human Genome. *Genome Biology and Evolution*, 5:606–620.

[8]. Zhao, Z., Guo, C., Sutharzan, S., Li, P., Echt, C. S., Zhang, J., & Liang, C. (2014). Genome-Wide Analysis of Tandem Repeats in Plants and Green Algae. *3G Genes/Genomes/Genetics*, 4: 67-78.

[9]. Dufresnes, C., Brelsford, A., Béziers, P., & Perrin, N. (2014). Stronger Transferability but Lower Variability in Transcriptomic- Than in Anonymous Microsatellites: Evidence from Hylid Frogs. *Molecular Ecology Resources*, 14:716–725.

[10]. Uygur Gocer, E., & Karaca M. (2016). Genetic Characterization of Some Commercial Cotton Varieties Using Td-DAMD-PCR Markers. *Journal of Scientific and Engineering Research*, 3:487-494.

[11]. Choi, K., Zhao, X., Kelly, K. A., Venn O., Higgins, J. D., Yelina, N. E., Hardcastle, T. J., Ziolkowski, P. A., Copenhaver, G. P., Franklin, F. C. H., McVean, G., &Hendersan, I. R. (2013). Arabidopsis Meiotic Crossover Hot Spots Overlap with H2A.Z Nucleosomes at Gene Promoters. *Nature Genetics*, 45:1327–1336.

[12]. Gao, C., Ren, X., Mason, A. S., Li, J., Wang, W., Xiao, M., & Fu, D. (2013). Revisiting an İmportant Component of Plant Genomes: Microsatellites. *Functional Plant Biology*, 40:645–645.

[13]. Zhao, H., Xing, Y., Liu, G., Chen, P., Zhao, X., Li, G., & Cai, L. (2015). GAA Triplet-Repeats Cause Nucleosome Depletion in the Human Genome. *Genomics*, 106:88–95.

[14]. Shilo, S., Melamed-Bessudo, C., Dorone, Y., Barkai, N., & Levy, A. A. (2015). DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in Arabidopsis. *Plant Cell*, 27:2427–2436.

[15]. Choi, K., & Henderson, I. R. (2015). Meiotic Recombination Hotspots – a Comparative View. *The Plant Journal*, 83:52–61.

[16]. Liu, C., Dou, Y., Guan, X., Fu, Q., Zhang, Z., Hu, Z., Zheng, J., Lu, Y., & Li, W. (2017). De novo Transcriptomic Analysis and Development of EST-SSRs for *Sorbus pohuashanensis* (Hance) Hedl. *PLoS ONE*, 12:e0179219.

[17]. Vukosavljev, M., Esselink, G. D., van'tWestende, W. P.C., Cox, P., Visser, R. G. F., Arens, P., & Smulders, M. J. M. (2015). Efficient Development of Highly Polymorphic Microsatellite Markers Based on Polymorphic Repeats in Transcriptome Sequences of Multiple İndividuals. *Molecular Ecology Resources*, 15:17–27.

[18]. Qu, J., & Liu, J. (2013). A Genome-Wide Analysis of Simple Sequence Repeats in Maize and the Development of Polymorphism Markers from Next Generation Sequence Data. *BMC Research Notes*, 6:403.

[19]. Karaca, M., & Ince, A. G. (2016). EpiOne: A Software Tool for Identification of Potential Cytosine DNA Methylation Marks in Promoters and Gene Bodies. *Journal of Scientific and Engineering Research*, 3:295-30.

[20]. Bilgen, M., Karaca, M., Onus, A. N., & Ince A. G. (2004). A Software Program Combining Sequence Motif Searches with Keywords for Finding Repeats Containing DNA Sequences. *Bioinformatics*, 20:3379-3386.

[21]. Karaca, M., Bilgen, M., Onus, A. N., Ince, A. G. & Elmasulu, S. Y. (2005). Exact Tandem Repeats Analyzer (e-TRA) for DNA Sequence Mining. *Journal of Genetics*, 84: 49-54.

[22]. Lawson, M. & J., Zhang, L. (2008). Housekeeping and Tissue-Specific Genes Differ in Simple Sequence Repeats in the 5′UTR Region. *Gene*, 407:54–62.

[23]. Zhou, Y., Liu, J., Han, L., Li, Z.-G., & Zhang, Z. (2011). Comprehensive Analysis of Tandem Amino Acid Repeats from Ten Angiosperm Genomes. *BMC Genomics*, 12:632.

[24]. Joshi-Saha, A., & Reddy, K. S. (2015). Repeat Length Variation in the 5'UTR of Myo-İnositol Monophosphatase Gene is Related to Phytic Acid Content and Contributes to Drought Tolerance in Chickpea (*Cicerarietinum* L.). *Journal of Experimental Botany*, 66:5683–5690.

[25]. Kumar, S., & Bhatia, S. (2016). A polymorphic (GA/CT)n- SSR İnfluences Promoter Activity of Tryptophan Decarboxylase Gene in *Catharanthus roseus* L. Don. *Scientific Reports*, 6:33280.

[26]. Zhang, L., Zuo, K., Zhang, F., Cao, Y., J., Wang, J., Zhang, Y., Sun, X., & Tang, K. (2006). Conservation of Noncoding Microsatellites in Plants: İmplication for Gene Regulation. *BMC Genomics*, 7: 323.

[27]. Nalavade, R., Griesche, N., Ryan, D. P., Hildebrand, S., & Krau, S. (2013). Mechanisms of RNA-İnduced Toxicity in CAG Repeat Disorders. *Cell Death Disease*, 4:e752.

[28]. Taka, S., Gazouli, M., Politis, P. K., Pappa, K. I., & Anagnou, N. P. (2013). Transcription Factor ATF-3 Regulates Allele Variation Phenotypes of the Human SLC11A1 Gene. *Molecular Biology Reports*, 40:2263–2271.

[29]. Quilez, J., Guilmatre, A., Garg, P., Highnam, G., Gymrek, M., Erlich, Y., Joshi, R. S., Mittelman, D., & Sharp, A. J. (2016). Polymorphic Tandem Repeats within Gene Promoters Act as Modifiers of Gene Expression and DNA Methylation in Humans. *Nucleic Acids Research*, 44:3750–3762.

[30]. Fuda, N. J., Guartin, M. J., Sharma, S., Danko, C. G., Martis, A. L., Siepel, A., & Lis, J. T. (2015). GAGA Factormaintains Nucleosome-Free Regions and has a Role in RNA Polymerase II Recruitment to Promoters. *PLoS Genetics*, 11:e1005108.

[31]. Chen, H. Y., Ma, S. L., Huang, W., Ji, L., Leung, V. H. K., Jiang, H., Yao, X.,& Tang, N. L. S. (2016). The Mechanism of Transactivation Regulation due to Polymorphic Short Tandem Repeats (STRs) Using IGF1 Promoter as a Model. *Scientific Reports*, 6:38225.

[32]. Hansel-Hertsch, R., Beraldi, D., Lensing, S. V., Marsico, G., Zyner, K., Parry, A., Di Antonia, M., Pike, j., Kimura, H., Narita, M., Tannahill, D., &Balasubramanian, S. (2016). G-quadruplex Structures Mark Human Regulatory Chromatin. *Nature Genetics*, 48:1267–1272.

[33]. Li, Y-C., Korol, A. B., Fahima, T., & Nevo, E. (2004). Microsatellites within Genes: Structure, Function, and Evolution. *Molecular Biology and Evolution*, 21:991-1007.

[34]. Kotrappa, N., Hendre, P. S., & Rathinavelu, R. (2017) Intra and Intergeneric Transferable Gene-Derived Orthologous Microsatellite Markers in Eucalyptus and Corymbia species. *Journal of Forest Research*, 22:65-68.

[35]. Sawaya, S. M., Lennon, D., Buschiazzo, E., Gemmell, N., &Minin, V. N. (2012). Measuring Microsatellite Conservation in Mammalian Evolution with a Phylogenetic Birth–Death Model. *Genome Biology and Evolution*, 4:636–647.

[36]. Press, M. O., Carlson, K. D., & Queitsch, C. (2014). The Overdue Promise of Short Tandem Repeat Variation for Heritability. *Trends Genetics*, 30:504–512.

[37]. Ribeiro, M. M., Teixeira, G. S., Martins, L., Marques, M. R., de Souza, A. P., & Line, S. R. (2015). G-quadruplex Formation Enhances Splicing Efficiency of PAX9 İntron 1. *Human Genetics*, 134:37–44.

[38]. Loire, E., Higuet, D., Netter, P., &Achaz, G. (2013). Evolution of Coding Microsatellitesin Primate Genomes. *Genome Biology and Evolution*, 5:283–295.

[39]. Schaper, E., Gascuel, O., & Anisimova, M. (2014). Deep Conservation of Human Protein Tandem Repeats within the Eukaryotes. Molecular Biology and Evolution. 31:1132–1148.

[40]. Sjakste, T., Paramonova, N., Sjakste, N. (2016). Structural and Functional Significance of Microsatellites. *Biopolymers and Cell*, 32:334–346.

[41]. Gymrek, M. (2017). A Genomic View of Short Tandem Repeats. *Current Opinion Genetics & Development*, 44:9–16.

[42]. Haasl, R. J., & Payseur, B. A. (2014). Remarkable Selective Constraints on Exonic Dinucleotide Repeats. *Evolution*, 68:2737–2744.

[43]. Amiteye, S., Corral, J.M., Vogel, H., Kuhlmann, M., Mette, M.F., & Sharbel, T.F. 2013. Novel MicroRNAs and Microsatellite-like Small RNAs in Sexual and Apomictic Boechera species. *MicroRNA*, 2:45-62.