Electronic Journal of Biotechnology 50 (2021) 37–44

Contents lists available at ScienceDirect

Electronic Journal of Biotechnology

journal homepage:

Research Article

Short Tandem Repeats in plants: Genomic distribution and function prediction

Lin Zhu [a,1], Hui Wu [b,1], Huayang Li [b], Heng Tang [b], Li Zhang [b], Houjuan Xu [b], Fangchan Jiao [c], Nian Wang [a,*], Long Yang [b,*]

[a] State Key Laboratory of Crop Biology, Shandong Agricultural University, Tai'an 271000, China
[b] Agricultural Big-Data Research Center and College of Plant Protection, Shandong Agricultural University, Taian 271018, China
[c] Yunnan Academy of Tobacco Agricultural Sciences, Kunming 650021, Yunnan, China

ARTICLE INFO

Article history:
Received 31 July 2020
Accepted 21 December 2020
Available online 30 December 2020

Keywords:
Algae
Coding sequences
Dinucleotide
GC contents
Genome
Hexanucleotide
Plants
STR density
Tandem repeats
Trinucleotide

ABSTRACT

Background: Short Tandem repeats (STRs) existed as popular elements in both eukaryotic and prokaryotic genomes.
Results: In this study, we analyzed the characteristics, distributions, and motif features of STRs within whole-genomes of 140 plant species. The results showed that STR density was negatively correlated with the genome size. Hexanucleotide repeat was the most abundant type of STRs. The distribution of algae shows a preference different from that of other plants. By analyzing GC contents of STRs and genome, it was concluded that STR motif was influenced by GC contents. Analysis of the long STRs in genome (length $\geq$ 1000 bp) found that dicots have the more long STRs. For STR types, di- and tri-nucleotide accounted for the highest proportion. Analyzing and designing long STRs in CDS (length $\geq$ 500 bp) was to verify the role of long STRs in *Gossypium hirsutum* TM-1 and *Solanum tuberosum*. By comparing the long STRs found in *Fragaria x ananassa* with other species, some evolutionary characteristics of the long STRs were obtained.
Conclusions: We got the characteristics, distribution, and motif features of STRs in the whole genome of 140 plants and obtained some evolutionary characteristics of long STRs. The study provides useful insights into STR preference, characteristics, and distribution in plants.
How to cite: Zhu L, Wu H, Li H, et al., Short Tandem Repeats in plants: genomic distribution and function prediction. Electron J Biotechnol 2021;50.https://doi.org/10.1016/j.ejbt.2020.12.003

1. Introduction

Tandem repeats comprising microsatellite DNA (short tandem repeats, STRs, 1–10 bp repeat unit size), minisatellite DNA (10–100 bp repeat unit size), and longer satellite DNA (repeat units of >100 bp) are distributed in plant genomes [1], especially in the euchromatin of eukaryotes, coding and noncoding nuclear and organellar DNA [2,3]. The number of tandem repeats is variable due to susceptibility to slippage events during DNA replication [4]. STRs, especially microsatellite markers have been commonly used in population genetics [5] due to properties, such as high mutation rate, codominance, reproducibility, and easy transferability among closely related species [6,7].

STRs have different motif types of preference among distantly related species [8,9] or even among closely related species [10,11]. STR preference refers to the different proportions of STR motifs in plants. The composition of STRs reflects the preference of STRs in "types of STRs," "GC content," "Length," and other aspects. STRs are mainly present within noncoding regions [8]. The STR motif types are affected by GC content [1,12] with more GC-rich motifs being present within higher GC content regions. GC-rich STRs tend to form secondary structures, such as a loop or a hairpin, which is formed by unstable trinucleotide repeats [13]. Alterations in these secondary structures may alter the expression of adjacent genes.

The presence of long STRs in CDS could enrich one or several amino acids, which may be used to alleviate abiotic stresses in crop plants [14], affect growth, yield, and cadmium uptake in *Triticum aestivum* L [15]. A 161 bp long STRs in centromere has been

Peer review under responsibility of Pontificia Universidad Católica de Valparaíso
* Corresponding authors.
E-mail addresses: nian.wang@sdau.edu.cn (N. Wang), lyang@sdau.edu.cn (L. Yang).
[1] These authors contributed equally to the publication.

https://doi.org/10.1016/j.ejbt.2020.12.003

confirmed to play a role in rapid evolution of *Genlisea* genome [16]. The repeat proteins are assumed to facilitate transient interactions amidst proteins and associated functional domains allow them to play key roles in diverse cellular processes conferring adaptation and tolerance to various environmental conditions in plants [17].

Recent STR studies mainly focused on the regulation of human diseases [18] and gene expression [19,20] and mainly on molecular markers in plants rather than on the difference in STR types among plants. Genome-wide analysis of STRs among many species is facilitated by recent advances in next-generation sequencing technologies [21]. In this study, 140 plants including algae, monocots, dicots, mosses, ferns, and gymnosperms were used to characterize STR types and functions of the special long STRs. The correlation between STRs and species evolution and kinship found in this study will enhance our understanding of STRs and their biological functions in plants. More importantly, this work provides useful insights into the structure and distribution characteristics of STRs as well as their variation patterns among different species.

2. Materials and methods

2.1. Sources of genome and CDS sequences

Genome information was obtained based on databases and the published literature. The complete genome sequences of 140 plants were downloaded from the public platforms. The data were divided into two types: genomic sequences and coding sequences (CDS). Some CDS were obtained from databases that provide the genome, while others are extracted based on custom Perl scripts. These plants are categorized into six groups: algae, mosses, ferns, gymnosperms, dicots, and monocots (Table S1).

2.2. STR detection and analysis

MISA (MIcroSAtellite identification tool; http://pgrc.ipk-gatersleben.de/misa/misa.html) was used to identify STRs with perfect tandem repeats and compound tandem repeats [22]. The minimum number of repeats was set as 12 for mono-nucleotide, six for di-, four for tri-, three for tetra- and penta-, and two for hexa- to deca-nucleotide repeats. In addition, STR motifs and their corresponding reverse complement motifs (e.g., CAG and GTC) were analyzed, respectively. Custom Perl scripts were used to initiate STR detection and to parse results from MISA. Only valid nucleotides (A, T, G, and C) were counted when analyzing the sequences.

The number and density of STRs were calculated with STR density being expressed as the length of detected STRs out of the total length of the sequences (bp/Mbp). GC contents were calculated for the six taxonomic groups using custom Perl scripts. The GC content of STRs were calculated in genome (e.g., GC content of $(ACTG)_3$ was $(2 \times 3) \div (4 \times 3) \times 100\%$). Pearson correlation (r) test was carried out in SPSS 25.0 (https://www.ibm.com/analytics/spss-statistics-software). Plots were made using ggplot2 of R and figures were used R 3.5.3 [23].

2.3. Long STR analysis and function predicted

Perl scripts were used to extract CDS and determine long STRs. Identified STRs within CDS were analyzed using custom Perl scripts. BLASTN (Basic Local Alignment Search Tool of nucleotide) was through a local Linux server. Alignment sequence determines the parameter as *E*-value $<10^{-5}$. Long STRs with a length greater than 500 bp extracted from CDS were analyzed to study its functional role. Their amino acid sequence was translated by NCBI ORF (open reading frame) finder (https://www.ncbi.nlm.nih.gov/orffinder/) with "ATG" setting as the ORF starting codon only. Protein secondary structure was predicted using PSIPRED 4.0 (http://bioinf.cs.ucl.ac.uk/psipred/). Online BLASTP (Basic Local Alignment Search Tool of Protein, https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to investigate the role of long repeating sequences. Alignment sequence determines the parameter as query cover >70%, *E*-value $<10^{-20}$, and present identity >80%. Abnegated function annotations are "predicted" and "hypothetical." The prediction of conservative structural domain used was NCBI CD-search (http://teachline.ls.huji.ac.il/72682/tutorials/RTS/ns4b-CD-Search.htm). Only sequences with "specific hits" were selected. Long STRs of *Gossypium hirsutum* and *Solanum tuberosum* were designed as 250 bp and 200 bp repeats decreasing to predict their effects on protein structure.

3. Results

3.1. STR distribution

A total of 283,867,588 STRs (length $\geq$ 12 bp) were identified from the 140 plant species, and their distribution patterns were characterized in the six taxonomic groups. As expected, the STR abundance is correlated with genome size (GS) (r = 0.977, $P < 0.01$, Fig. 1A). Unlike STR abundance, there is a negative correlation ($r = -0.305$, $P < 0.01$, Fig. 1B) between STR density (i.e., bp covered by STRs per Mb of the genome) and the genome size. STR density varies considerably among plants with small GS; however, STR density tends to decrease in plants with large GS (Fig. 1B).

STR density ranges from 9.3 kb/Mb in *Larix sibirica* to 58 kb/Mb in *Micractinium conductrix*. The maximum density of algae STRs is six times bigger than the minimum (Fig. 2). STR density is not significantly different among the three mosses, and the density of *Physcomitrella patens* is 1.5 times that of *Marchantia polymorpha*. STR density varies 2.4-fold among the three fern species of the same genus. Among the nine gymnosperm species, the highest STR density is only 19.4 kb/Mb and the average is 15.1 kb/Mb. Among monocots, the STR coverage of Gramineae ranged from 1.2% to 5.2%, while that of Orchidaceae, ranging from 1.9% to 2.9%. The STR density in dicots ranges from 14 kb/Mb to 35 kb/Mb with an average of 23 kb/Mb.

3.2. Distribution of STR types and motif preference

The GC content of the STRs is highly correlated with the genomic GC content (r = 0.947, $p < 0.01$, Fig. 3). For example, the genomic GC content of algae ranges from 36.03% to 65.68% and the GC content of STRs ranges from 27.54% to 87.32%.

To analyze STR motif preference, we counted the top ten most abundant STR motifs in each species. As expected, species with a higher GC content usually has more GC-rich STR motif (Table S2) with the exception of some species of Poaceae. For example, in *Sorghum bicolor* genome, despite the fact that most of its motifs are GC-rich, the GC content of STRs is 40.02%. GC/CG motif ranks top ten only in algae, but not in other plants.

In terms of repetition types, in the top 10 statistics, mononucleotide to hexanucleotide occurred except for pentanucleotide. Only *Sorghum bicolor* (Poaceae) showed heptanucleotide, and *Vigna angularis* (Fabaceae) showed nonanucleotide. The STR motif distribution patterns of algae were diverse, The STR motif distribution patterns of dicots were mainly AT rich, and the motif types were generally A/T and AT/TA. Closely related species showed similar types of motifs, and motif preference in the three mosses was similar to that of dicots. STR GC content of *Physcomitrella patens* was only 17.59%, while AT-rich motifs did not dominate among the top10 motifs (Table S2). For the three ferns, their motif distribution is very different from that of dicots. For gymnosperms, *Ginkgo*

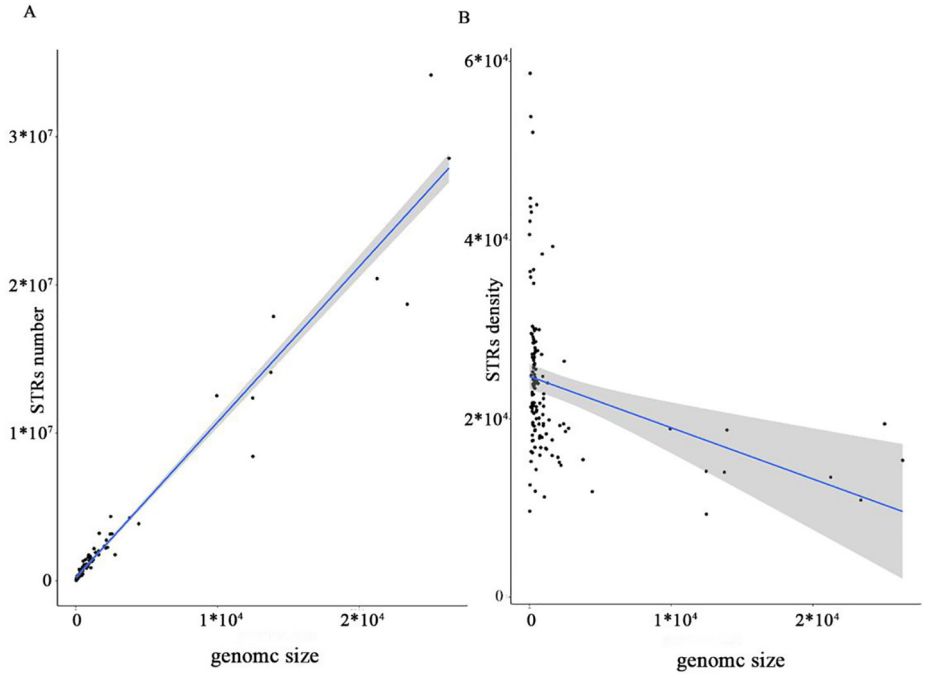

Fig. 1. (A) Relationship between genome size with STR number; (B) relationship between genome size with STR density.

biloba and *Pinus taeda* showed little difference in motif distribution. Among monocots, the motif of GC-rich in Poaceae is more common than that in Orchidaceae.

3.3. Genomic patterns of STRs categorized by motif size

STRs with hexanucleotides were the most abundant type in the six groups, ranging from 30% to 64%. The frequency of heptanucleotide repeats is second to that of hexanucleotides except for algae with trinucleotide repeats second to most. The least abundant type of repeat was decanucleotide accounting for between 0.6% and 2% (Fig. 4).

3.4. The preference of long STR motif

A total of 2025 long STRs (with a length ≥1000 bp) were identified in genomes of 58 plants. A statistical of the occurrence of long repeating sequences increasing every 1000 bp were analyzed, with a large proportion of dicots in each gradient. In order to get rid of the statistical effect of the large proportion of dicots, we adopted the algorithm of statistical species/total species (Fig. 5A). Dicots appeared in all gradient statistics, while mosses and ferns did not. In the statistics of STR length greater than 10,000 bp, a total of 22 STR sequences were identified from 12 species: one gymnosperm (*Ginkgo biloba*), three monocots (*Apostasia shenzhenica*, *Gastrodia elata, and Phalaenopsis equestris*) and eight dicots (*Brassica napus, Citrus medica, Citrus_sinensis, Gossypium_arboreum, Gossypium_barbadense, Medicago truncatula, Nicotiana sylvestris,* and *Solanum pennellii*).

Dinucleotide and trinucleotide account for a large proportion of all SSR types (Fig. 5B). Repeat types from mono- to deca-nucleotide all appeared in long repeat sequence with a length between 1000 and 5000 bp. However, in long repeat sequence with a length above 5000 bp, some STR types were absent. Compound TR, mono- and di-nucleotide were present in all long repeat sequences with different lengths, while only trinucleotide was absent in repeat sequences with a length between 7000 and 8000 bp. The most missing types were penta-, octa-, nona- and deca-nucleotide. In long repeated sequences over 10,000 bp, repeat type focused on 1–6 bp, in which one mononucleotide, five dinucleotide, nine trinucleotide, one tetranucleotide, and two hexanucleotide. Motif type is dominated by poly-A/T, and only one sequence was based on poly-C. Through the analysis of all composite STRs that appeared, it was found that one (or several) motifs of most composite STRs originated from mutations in other motif bases.

3.5. Long STRs of CDS analysis and protein structure prediction

Cultivated strawberry (*Fragaria x ananassa*) has the highest number of long STRs (>500 bp) with a total of 53 gene sequences detected. The CDS sequences of *Fragaria vesca*, *Fragaria iinumae*, *Fragaria nipponica*, *Fragaria nubicola,* and *Fragaria orientalis* were used to BLAST with these gene sequences, and resulted in 46, 282, 285, 212, and 275 hits, respectively. Type similarity is very high and the difference is only the number of repetitions in comparing long STRs with the TR types in these similar sequences.

Ninety-nine long STRs (>500 bp) were detected in 15 species with trinucleotide and hexanucleotide accounting for 64% and 8%, respectively. Ten long STRs were not translated into amino acids and 14 sequences containing the conserved domain were obtained whether repeat sequences are removed (Fig. 6). BLASTP results yielded 25 sequences with the motif GAA/AGA/AAG appeared in 16 proteins, of the amino acids translated, 12 sequences were repeated with glutamate, and four were lysine repeats. In the translated amino acid chain, glutamate and lysine repeats account for 27.8% and 12.2%, respectively. CD-search and BLASTP were performed for the protein sequences after decreasing 200 bp repetitive sequence treatment, but no functional domains were found.

In *G. hirsutum*, "AAG" is the motif of long STRs within three gene sequences, which have a length of 855 bp, 834 bp, and 813 bp, respectively. In *S. tuberosum*, four genes that contain the long STRs varied from 567 bp to 605 bp in length, and the STR motif is not uniform. When the length of the gene sequence containing the long STRs was reduced, a series of changes occurred in the protein

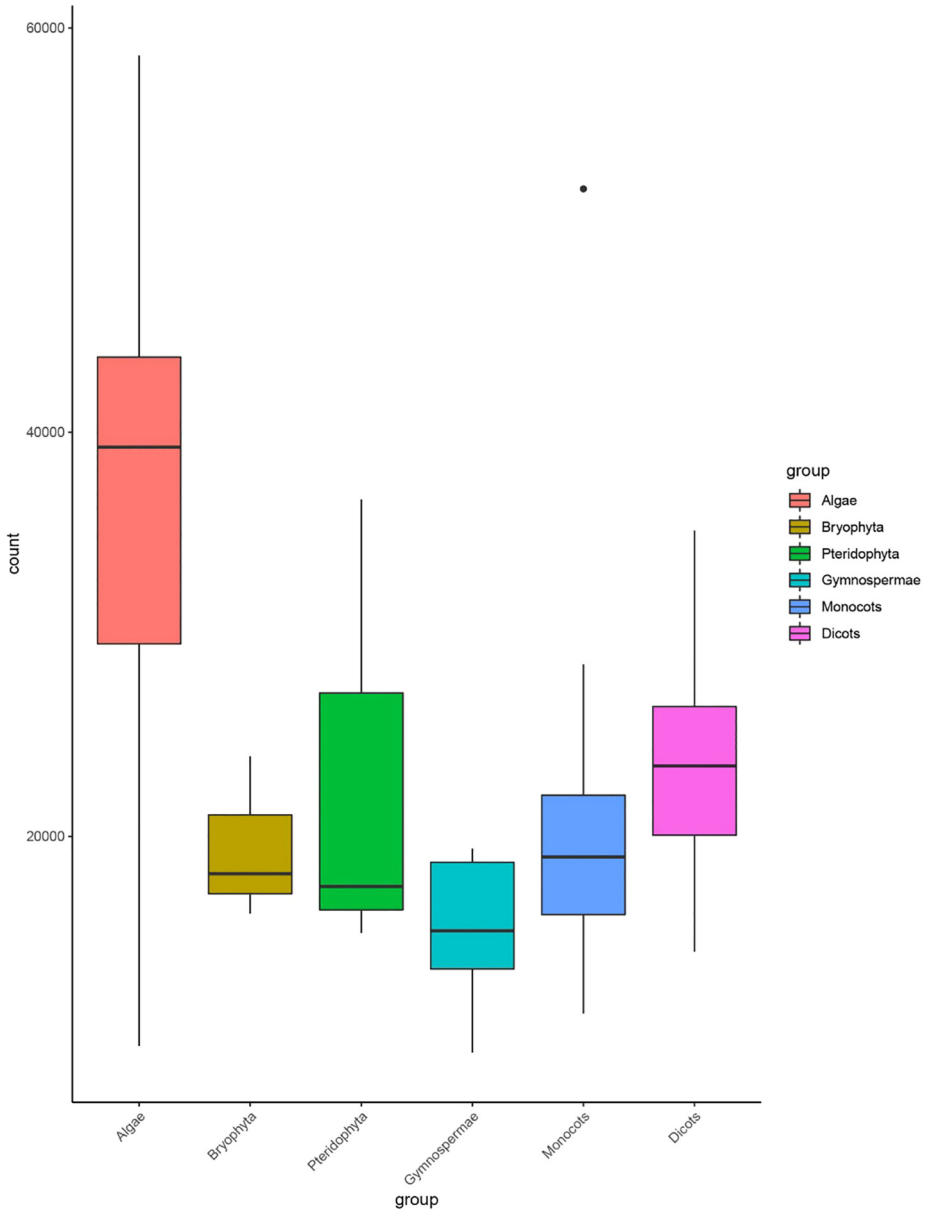

Fig. 2. STR density distribution of six groups.

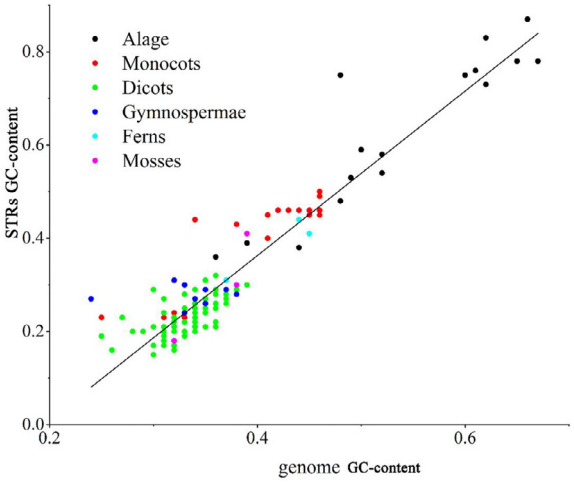

Fig. 3. The relationship with GC content in STR and genome.

sequence. The changes in the repetitive sequence caused them to change the number of amino acids and the position of the starting codon in ORF, making the translated protein sequence different. Secondary structure prediction showed α-helix and the random coil were the main structures of the protein sequence containing the long STRs. Only one sequence of *S. tuberosum* has the structure of cytoplasmic proteins, accounting for 10.71%. Helical accounts for over 90% in the three sequences containing long STRs in upland cotton. In *S. tuberosum*, main structures of the four genes containing long STRs are irregular coil for ST_400058467, flat-helix for ST_400068108, α-helix for ST_400039585 and membrane protein structure ST_400022514, accounting for 88.98%, 59.89%, 45.8%, and 10.71%, respectively. After the processing, there was an obvious structural change in the sequence. The n-fold appears in gh_cotad_07805-250, and the n-fold disappears in st_400058467-200. The membrane protein structure of st_400022514-200 and st_400022514-400 disappeared. The repeated motif of three upland cotton and potato ST_400022514 is GAA, which is a triplet motif. The subsequent length reduction has little effect on the

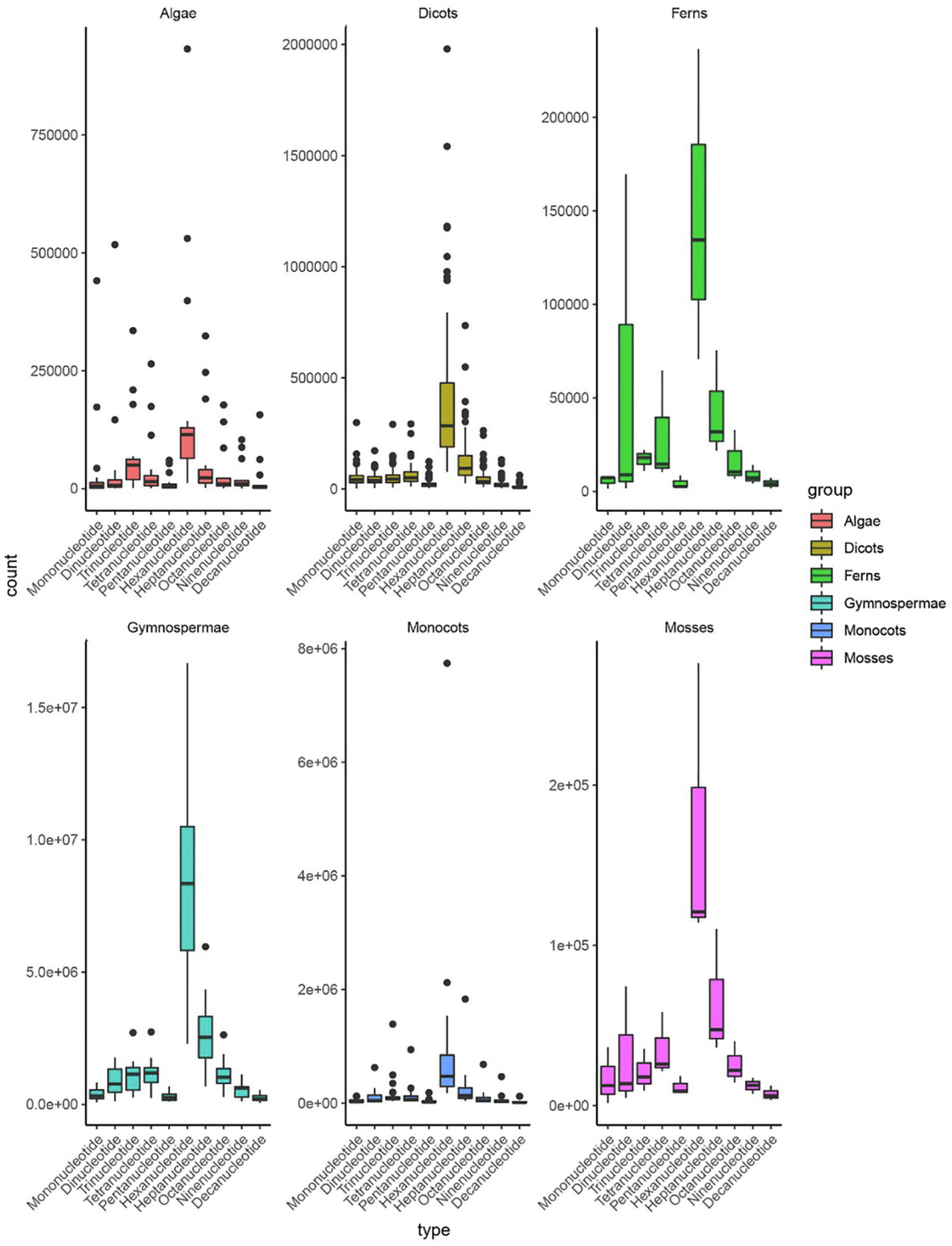

Fig. 4. STR types in six groups.

structure. In comparison, the motif of the other three long STR sequences in *S. tuberosum* is non-triplet repetition, and the length change has a greater impact on the subsequent amino acids and secondary structure of proteins (Table 1).

4. Discussion

Tandem repeats are supposed to contribute to large genome size. Large genome size provides more chances for the production of STRs. However, STR density was weakly correlated with genome size which was possibly due to the underrepresentation of STRs in the repetitive elements of the plant genome that involved in genome expansion [24]. STR density in monocots was higher than that in dicots [25]. Algae show wide variation in relative abundance and density of STRs compared with flowering plants. Similar results were found in the chloroplast genome of plants [26].

STR distribution is biased, which suggests that different STRs play different roles in different genic regions [27]. STRs also have functional importance in regulating gene expression [19,20]. STRs in eukaryotes are more common in noncoding regions. poly-A/T was more abundant than poly-C/G as for mononucleotide in the

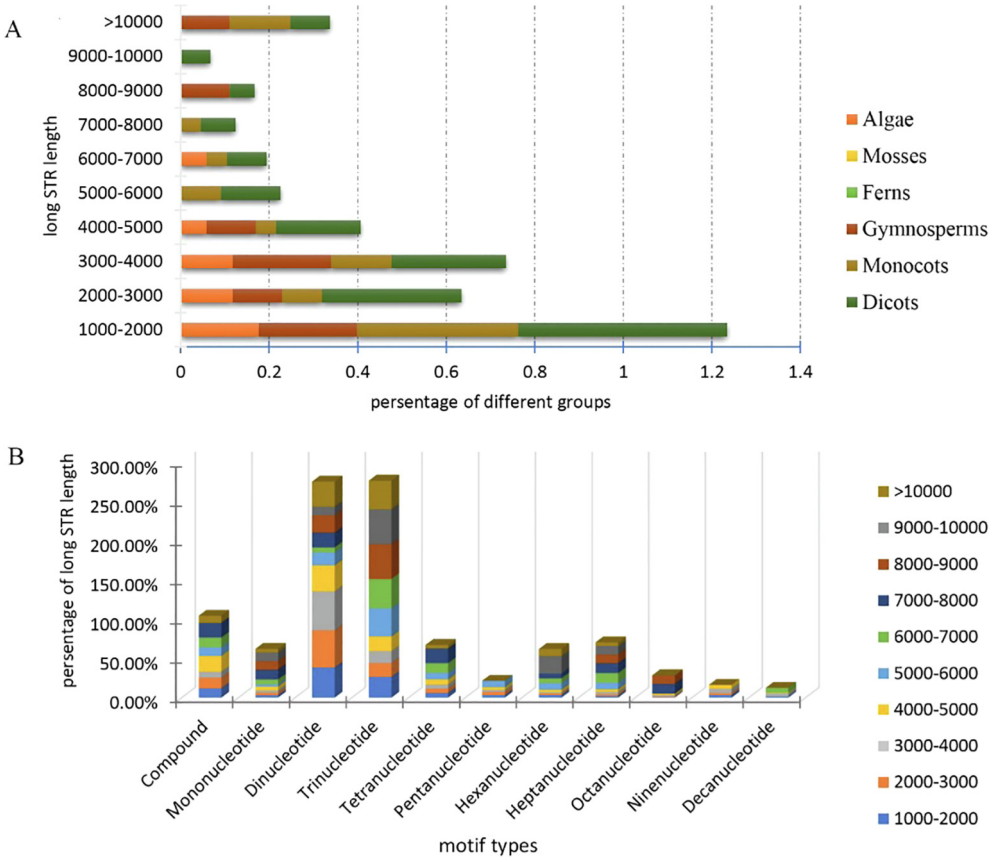

Fig. 5. (A) Species with long repeating sequences in six groups; (B) the proportion of STR types in each gradient.

studied plant species, consistent with earlier reports [10,11]. Recent studies have shown that poly-A/T STRs are associated with meiosis recombination hotspots [28]. Advanced plants tend to be zoogamy in nature. This may explain why AT-rich motif in seed plants are common than algae.

STRs are more likely to be produced in genomes with a high or low GC content [29]. The GC content of genomes affects the density of GC-rich repeats. High frequency of GC-rich repeats usually occurs in genome with a high GC content [30]. GC content in Poaceae and some algae is higher than that in other species, possibly due to gene structure, recombination patterns, and GC-biased gene conversion [31]. Genomic adaptations associated with changing GC content may play an important role in the evolution of the Earth's contemporary biota, such as the rise of grass-dominated biomes during the mid-tertiary [32]. One of the major selective advantages of GC-rich DNA is hypothesized to facilitate more complex gene regulation.

Long STRs mainly have the di- and tri-nucleotide types, probably because short repeats are the starting point for subsequent extensions of STRs [33], which are the most variable type of DNA sequence in the genome [34]. For prokaryote, long STRs unusually act as translation and transcriptional "switches." Their presence is maintained by positive selection [35]. In the predicted 14 long repeat sequences containing conservative structural domains, 11 repeat sequences have a motif of trinucleotide. The remaining three long repeat sequences were multiplication of trinucleotide, with conservative domain in the upstream of the repeat sequence. Different motifs translate into the same amino acids, which may be related to the position of repeated sequences read by codons. Repeat unit is a three-fold nucleotide to avoid frame-shift mutations that affect translation [36]. As species evolved and biological complexity increased, more STR motifs had accumulated long repeats, especially in dicots. This suggests that certain motifs have been selected for potential functional purposes.

Emergence of repeated domains seems to be evolutionarily favored due to their involvement in plant-specific key functional processes such as transcription, protein binding, hormone, and secondary metabolite pathways [17]. Tandem duplications are an integral part of protein evolution. The enrichment of amino acids may also lead to pathological changes in organisms, such as the CAG triad in humans, which represents a group of neurodegenerative diseases that are related to polyglutamine repeat domain amplification [37]. Commonly occurring repeat proteins across eukaryotic genome are enriched with imperfect STRs. However, perfect and nearly perfect repeats also occur in a large number in genome. Different mutations such as insertions, substitutions, deletions contribute to sequence imperfection within repeats. The change of repeat sequence not only leads to the change of amino acid, but also leads to the change of protein structure, and finally results in the loss of protein function.

Conflict of interest

The authors declare no conflict of interest.

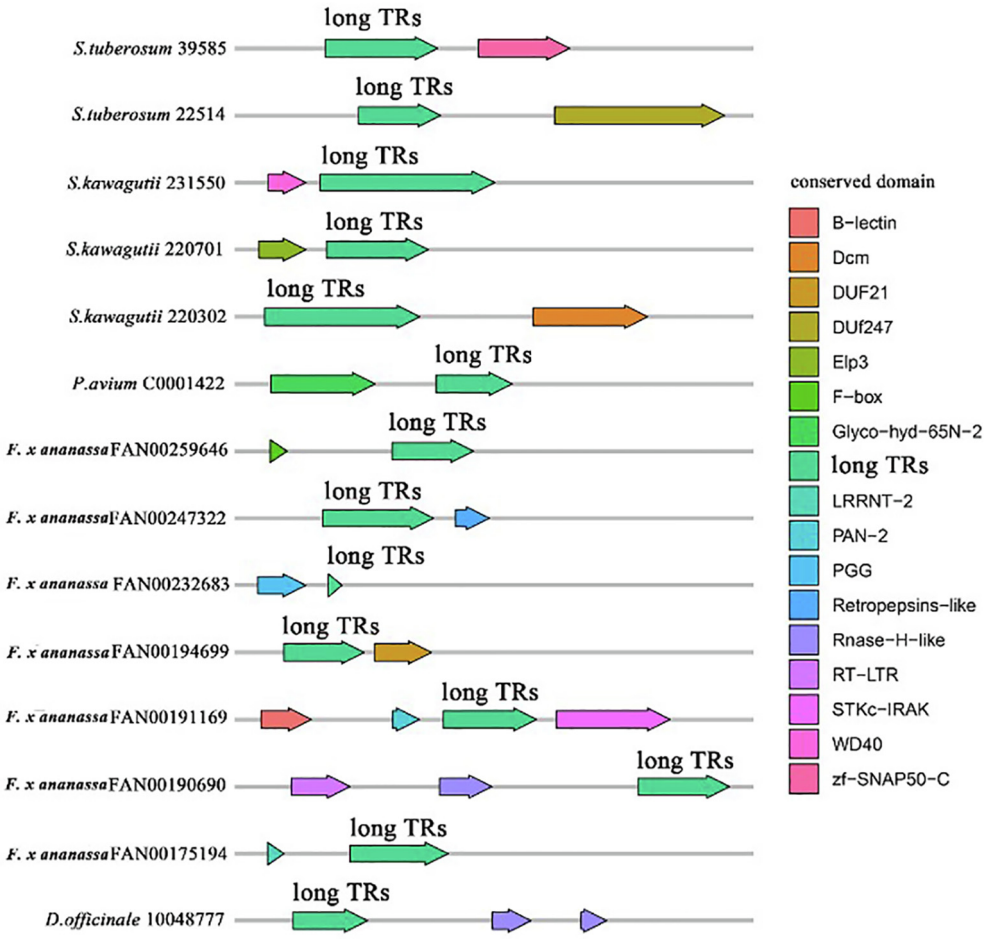

Fig. 6. The relationship between the length and position of the conserved domain and long STRs in the coding region sequence. On the left is the species name and gene id, in the middle is the location of the conserved domain and long STRs, and on the right is the conserved domain type.

Table 1
Statistical results of gene sequences with long STRs in upland *Gossypium hirsutum* and *Solanum tuberosum* after degressive treatment.

CDS-ID (designed)	Repeat Motif	Amino Acids Repeats	Total	β-Strand	Proportion (%)	α-helix	Proportion (%)	Coil	Proportion (%)	cytoplasmic	Proportion (%)
GH_CoTAD_07805	GAA	E	295	0	0	274	92.88	21	7.12	0	0
GH_CoTAD_07805-250		K	212	3	1.42	185	87.26	24	11.32	0	0
GH_CoTAD_07805-500		R	129	0	0	111	86.05	18	13.95	0	0
GH_CoTAD_54284	GAA	K	295	0	0	291	98.64	4	1.36	0	0
GH_CoTAD_54284-250		R	212	0	0	198	93.4	14	6.6	0	0
GH_CoTAD_54284-500		E	129	0	0	125	96.9	4	3.1	0	0
GH_CoTAD_65565	GAA	E	301	0	0	292	97.01	9	2.99	0	0
GH_CoTAD_65565-250		K	209	0	0	202	96.65	7	3.35	0	0
GH_CoTAD_65565-500		E	151	0	0	142	94.04	9	5.96	0	0
ST_400022514	GAA	E	1120	46	4.11	513	45.8	441	39.38	120	10.71
ST_400022514-200		R	859	38	4.42	464	54.02	357	41.56	0	0
ST_400022514-400		R	802	50	6.23	445	55.49	307	38.28	0	0
ST_400039585	CTCTCGT	LVSRLLS	743	67	9.02	445	59.89	231	31.09	0	0
ST_400039585-200		LVSRLLS	387	2	0.52	302	78.04	83	21.45	0	0
ST_400039585-400		SRLLSLV	323	2	0.62	233	72.14	88	27.24	0	0
ST_400058467	CAACT	TQLNS	224	5	2.23	0	0	219	97.77	0	0
ST_400058467-200		QLNST	157	0	0	5	3.18	152	96.82	0	0
ST_400058467-400		QLNST	91	2	2.2	2	2.2	87	95.6	0	0
ST_400068108	GAGGAA	E	245	8	3.27	218	88.98	19	7.76	0	0
ST_400068108-200		GR	179	9	5.03	13	7.26	157	87.71	0	0
ST_400068108-400		KR	110	7	6.36	81	73.64	22	20	0	0

Financial support

This work was supported by the Foundation of Shandong Province Modern Agricultural Technology System Innovation Team (SDAIT-25-02) and the Foundation of Platform Construction for Tobacco Genome Breeding (2017YN05).

Supplementary material

https://doi.org/10.1016/j.ejbt.2020.12.003.

References

[1] Zhao Z, Guo C, Sutharzan S, et al. Genome-wide analysis of tandem repeats in plants and green algae. G3 (Bethesda) 2014;4:67–78. https://doi.org/10.1534/g3.113.008524.
[2] Pérez-Jiménez M, Besnard G, Dorado G, et al. Varietal tracing of virgin olive oils based on plastid DNA variation profiling. PLoS One 2013;8:. https://doi.org/10.1371/journal.pone.0070507. PMid: 23950947e70507.
[3] Phumichai C, Phumichai T, Wongkaew A. Novel chloroplast microsatellite (cpSSR) markers for genetic diversity assessment of cultivated and wild *Hevea* rubber. Plant Mol Biol Rep 2015;33:1486–98. https://doi.org/10.1007/s11105-014-0850-x.
[4] Richards RI, Sutherland GR. Simple repeat DNA is not replicated simply. Nat Genet 1994;6:114–6. https://doi.org/10.1038/ng0294-114. PMid: 8162063.
[5] Hu YN, Zhao L, Buggs RJ, et al. Population structure of Betula albosinensis and *Betula platyphylla*: evidence for hybridization and a cryptic lineage. Ann Bot 2019;123:1179–89. https://doi.org/10.1093/aob/mcz024. PMid: 30916314.
[6] Squirrell J, Hollingsworth PM, Woodhead M, et al. How much effort is required to isolate nuclear microsatellites from plants?. Mol Ecol 2003;12:1339–48. https://doi.org/10.1046/j.1365-294X.2003.01825.x. PMid: 12755865.
[7] Schlötterer C. The evolution of molecular markers — just a matter of fashion?. Nat Rev Genet 2004;5:63–9. https://doi.org/10.1038/nrg1249. PMid: 14666112.
[8] Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 2000;10:967–81. https://doi.org/10.1101/gr.10.7.967. PMid: 10899146.
[9] Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 2001;18:1161–7. https://doi.org/10.1093/oxfordjournals.molbev.a003903. PMid: 11420357.
[10] Xu Y, Hu Z, Wang C, et al. Characterization of perfect microsatellite based on genome-wide and chromosome level in Rhesus monkey (*Macaca mulatta*). Gene 2016;592:269–75. https://doi.org/10.1016/j.gene.2016.07.016. PMid: 27395431.
[11] Wang XT, Zhang YJ, Qiao L, et al. Comparative analyses of simple sequence repeats (SSRs) in 23 mosquito species genomes: Identification, characterization and distribution (Diptera: Culicidae). Insect Sci 2019;26:607–19. https://doi.org/10.1111/1744-7917.12577. PMid: 29484820.
[12] Victoria FC, Maia LC, Oliveira AC. *In silico* comparative analysis of SSR markers in plants. BMC Plant Biol 2011;11:15. https://doi.org/10.1186/1471-2229-11-15. PMid: 21247422.
[13] Lenzmeier BA, Freudenreich CH. Trinucleotide repeat instability: a hairpin curve at the crossroads of replication, recombination, and repair. Cytogenet Genome Res 2003;100:7–24. https://doi.org/10.1159/000072836. PMid: 14526162.
[14] Sadak MS, Abdelhamid MT. Influence of amino acids mixture application on some biochemical aspects, antioxidant enzymes and endogenous polyamines of *Vicia faba* plant grown under seawater salinity stress. Gesunde Pflanzen 2015;67:119–29. https://doi.org/10.1007/s10343-015-0344-2.
[15] Rizwan M, Ali S, Hussain A, et al. Effect of zinc-lysine on growth, yield and cadmium uptake in wheat (*Triticum aestivum* L.) and health risk assessment. Chemosphere 2017;187:35–42. https://doi.org/10.1016/j.chemosphere.2017.08.071. PMid: 28834770.
[16] Tran TD, Cao HX, Jovtchev G, et al. Centromere and telomere sequence alterations reflect the rapid genome evolution within the carnivorous plant genus *Genlisea*. Plant J 2015;84:1087–99. https://doi.org/10.1111/tpj.13058.
[17] Manisha S, Pandey GK. Expansion and function of repeat domain proteins during stress and development in plants. Front Plant Sci 2015;6:1218. https://doi.org/10.3389/fpls.2015.01218.
[18] Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet 2018;19:286–98. https://doi.org/10.1038/nrg.2017.115. PMid: 29398703.
[19] Gymrek M, Willems T, Guilmatre A, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet 2016;48:22–9. https://doi.org/10.1038/ng.3461. PMid: 26642241.
[20] Quilez J, Guilmatre A, Garg P, et al. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res 2016;44:3750–62. https://doi.org/10.1093/nar/gkw219. PMid: 27060133.
[21] Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet 2010;11:31–46. https://doi.org/10.1038/nrg2626.
[22] Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 2003;106:411–22. https://doi.org/10.1007/s00122-002-1031-0. PMid: 12589540.
[23] Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat 1996;5:299–314. https://doi.org/10.1080/10618600.1996.10474713.
[24] Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 2002;30:194–200. https://doi.org/10.1038/ng822. PMid: 11799393.
[25] Sonah H, Deshmukh RK, Sharma A, et al. Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. PLoS One 2011;6:e21298. https://doi.org/10.1371/journal.pone.0021298.
[26] George B, Bhatt BS, Awasthi M, et al. Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. Curr Genet 2015;61:665–77. https://doi.org/10.1007/s00294-015-0495-9. PMid: 25999216.
[27] Gao C, Ren X, Mason AS, et al. Revisiting an important component of plant genomes: microsatellites. Funct Plant Biol 2013;40:645. https://doi.org/10.1071/FP12325. PMid: 32481138.
[28] Heissl A, Betancourt AJ, Hermann P, et al. Length asymmetry and heterozygosity strongly influences the evolution of poly-A microsatellites at meiotic recombination hotspots. bioRxiv 2018:431841. https://doi.org/10.1101/431841.
[29] Tian X, Strassmann JE, Queller DC. Genome nucleotide composition shapes variation in simple sequence repeats. Mol Biol Evol 2011;28:899–909. https://doi.org/10.1093/molbev/msq266. PMid: 20943830.
[30] Jiang Q, Li Q, Yu H, et al. Genome-wide analysis of simple sequence repeats in marine animals-a comparative approach. Mar Biotechnol 2014;16:604–19. https://doi.org/10.1007/s10126-014-9580-1. PMid: 24939717.
[31] Glémin S, Clément Y, David J, et al. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. Trends Genet 2014;30:263–70. https://doi.org/10.1016/j.tig.2014.05.002. PMid: 24916172.
[32] Šmarda P, Bureš P, Horová L, et al. Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc Natl Acad Sci USA 2014;111:E4096–102. https://doi.org/10.1073/pnas.1321152111. PMid: 25225383.
[33] Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet 2004;5:435–45. https://doi.org/10.1038/nrg1348. PMid: 15153996.
[34] Weber JL. Informativeness of human $(dC\text{-}dA)_n{\cdot}(dG\text{-}d_T)n$ polymorphisms. Genomics 1990;7:524–30. https://doi.org/10.1016/0888-7543(90)90195-Z.
[35] Himmelreich R, Hilbert H, Plagens H, et al. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 1996;24:4420–49. https://doi.org/10.1093/nar/24.22.4420. PMid: 8948633.
[36] Legendre M, Pochet N, Pak T, et al. Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res 2007;17:1787–96. https://doi.org/10.1101/gr.6554007. PMid: 17978285.
[37] Verani M, Bustamante M, Martufi P, et al. Conformational modulation mediated by polyglutamine expansion in CAG repeat expansion disease-associated proteins. Biochem Biophys Res Commun 2016;478:949–55. https://doi.org/10.1016/j.bbrc.2016.08.057. PMid: 27520369.