Plant Omics Journal POJ

POJ 17(02):1-11 (2025) ISSN:1836-3644
doi: 10.21475/POJ.17.02.25.pne11

Research Note

Genome-scale mining of simple sequence repeats (SSRs) in the forage grass *Urochloa mosambicensis* using low-coverage whole-genome sequencing data: abundance, distribution, and composition

Ueslei Silva Leão[1, 2,*], Geice Ribeiro da Silva[2], Luíce Gomes Bueno[3], Aline Barbosa Negreiros[2], and Fábio Mendonça Diniz[4,*]

[1]Universidade Estadual do Piauí, UESPI, São Raimundo Nonato PI 64770-000, Brazil
[2]Northeast Biotechnology Network - RENORBIO/Animal Science Program - PPGCA, Universidade Federal do Piauí, Teresina PI 64049-550, Brazil
[3]Embrapa Gado de Corte, Avenida Rádio Maia, 830, Zona Rural, 79106-550, Campo Grande, MS, Brazil
[4]Embrapa Caprinos e Ovinos, Sobral CE 62010-970, Brazil

***Corresponding author: fabio.diniz@embrapa.br**

Abstract
Simple sequence repeats (SSRs), also known as microsatellites, are common components of prokaryotic and eukaryotic genomes. Microsatellite loci are widely applied as molecular marker systems in plant population studies and genetic breeding due to their codominant inheritance, high polymorphism, and reproducibility. The development of these markers, however, requires prior genomic information. Although low-coverage whole-genome sequencing using the Illumina MiSeq platform provides limited sequencing depth, it is usually sufficient to identify thousands of SSR regions. In this study, low-coverage sequencing was performed on the *Urochloa mosambicensis* genome using the Illumina MiSeq platform. This tropical forage grass shows strong potential for adaptability and persistence in dry environments, making it a promising feed source for ruminants in harsh conditions. The *U. mosambicensis* genome was screened for SSRs to evaluate their potential for molecular marker development. The high-quality Illumina sequencing reads were assembled into 32,931 contigs (N50 = 395 bp) using the CLC Genomics Workbench. The software Krait and MSDB were used to analyze the low-coverage *U. mosambicensis* sequencing data in order to identify and evaluate pure, interrupted pure, compound, and interrupted compound SSRs. A total of 2,665 pure microsatellites were identified, with the highest relative abundances found in trinucleotides (31.52 loci/Mb) and pentanucleotides (16.62 loci/Mb). Hexanucleotides (87.18%) and trinucleotides (86.96%) exhibited the highest proportion of sequences suitable for amplification. A greater abundance of interrupted compound motifs (1.81 loci/Mb) and pure compound SSRs (1.15 loci/Mb) was observed compared to interrupted pure SSRs (0.48 loci/Mb). This study also suggests that non-consensus bases positively influence the size of interrupted pure microsatellites, indicating greater stability against potential mutations. These findings provide a valuable resource for the selection of potential molecular markers for genetic breeding and population studies of the tropical forage species *U. mosambicensis*.

Keywords: Abundance, Microsatellites, *Capim-corrente*, Genetic breeding, Population studies, SSR.
Abbreviations: CDS_coding sequences; cSSR_compound simple sequence repeats; gDNA genomic DNA; lcWGS_low-coverage whole-genome sequencing; MSDB_Microsatellite Search and Building Database; NGS_next-generation sequencing; UTRs_untranslated regions.

Introduction

Simple sequence repeats (SSRs), also known as microsatellite loci, are common components of prokaryotic and eukaryotic genomes and can be found in both coding and non-coding regions. Microsatellites are tandemly repeated DNA sequences that consist of units ranging from 1 to 6 base pairs (mono-, di-, tri-, tetra-, penta-, or hexa-nucleotides) in length (Tautz, 1984; Chambers and MacAvoy, 2000). These types of repeated DNA sequences vary greatly across genomes of different species in terms of their relative abundance, frequency, density, and distribution (Weber, 1990; Toth et al., 2000; Katti et al., 2001). Due to their codominant inheritance, high polymorphism, characterized by considerable variation in the number of repeat units among individuals, and high reproducibility, these loci are widely used as molecular marker in plant genetics and breeding programs for tropical forage grasses. They are applied in diversity assessment (Garcia-Martinez et al., 2006), cultivar identification and protection (Ercisli et al., 2011; Zhang et al., 2016), molecular mapping (Danin-Poleg et al., 2000; Bindler et al., 2007), and marker-assisted selection (Sui et al., 2009).

Next-generation sequencing (NGS) technology stands out as one of the most promising approaches for the isolation and development of genomic and transcriptomic microsatellite markers in plants, as it generates thousands of fragments (reads) simultaneously and provides a large volume of sequence data for broader genomic coverage in a faster and simpler process (Zalapa et al., 2012; Taheri et al., 2018). The Illumina platform, in particular, has shown the capacity to deliver sufficient genome coverage at a lower cost compared to alternative techniques and eliminates the need for constructing microsatellite-enriched DNA libraries, which is a time-consuming and laborious procedure required by traditional methods (Abdelkrim et al., 2009; Ray and Satya, 2014).

A comprehensive analysis of a genome using low-coverage whole-genome sequencing (lcWGS) can offer valuable insights into the functional and evolutionary roles of repetitive sequences, helping to elucidate their origin and evolution (Toth et al., 2000; Katti et al., 2001; Lou et al., 2021; Chat et al., 2022), and enhancing their application as reliable molecular markers. Nevertheless, the wide variation in SSR abundance across forage species remains poorly understood.

Interest in *Urochloa* forage grasses has grown in recent years as rural producers recognize the benefits of these tropical grasses, particularly their productivity under harsh environmental conditions. For example, *Urochloa mosambicensis* (Hack.) Dandy, a species of the Poaceae family, demonstrates strong potential for adaptability and persistence in dry environments across various soil types, whether cultivated alone or in combination with other forage grasses (McIvor, 1984; Coates, 1997; Treydte et al., 2013). Despite its significant potential as a feed source for ruminants, this species remains poorly studied.

Significant efforts have recently been made to identify and develop SSR markers for various forage grasses in the *Urochloa* genus, including *U. brizantha* (Jungmann et al., 2009a), *U. ruziziensis* (Silva et al., 2013), *U. humidicola* (Jungmann et al., 2009b; Vigna et al., 2011; Santos et al., 2015), and *U. decumbens* (Ferreira et al., 2016), with a few microsatellites also developed more recently for *U. mosambicensis* (Leão et al., 2023).

In the present study, the *Urochloa mosambicensis* genome was sequenced at low coverage (lcWGS), mined for the presence of SSRs, and its potential for marker development was evaluated. In this context, an *in silico* analysis of microsatellite frequency and distribution was conducted, aiming to identify non-redundant SSRs and support their use in efficient marker development.

Results and Discussion

Low-coverage whole-genome sequencing

Low-coverage genome sequencing generated a total of 57,170,592 reads for *Urochloa mosambicensis*, which were assembled into 149,259 contig sequences. The N50 value was 395 bp, indicating that half of the assembled contigs are at least this length, which is sufficient to allow reliable primer design in the flanking regions immediately adjacent to the repeat motif, as required for successful marker development. Data quality assessment using FastQC during the preprocessing phase to eliminate redundancy resulted in a total of 32,931 contigs, with a cumulative length of 33,090,060 bp (33.09 Mb). The fragment sizes ranged from 94 to 35,547 bp, with an average length of 1,005 bp and a GC content of 47.11%. This GC content is higher than that reported for the grass *Setaria italica*, which has a GC content of 45.6% (Deng et al., 2016). Sequences containing simple sequence repeats were deposited in GenBank under accession numbers OP806325-OP806384 and MH742936-MH742955.

Overall, the relative abundance detected in this study (80.54 loci/Mb) was lower than that observed in microsatellite studies of different species in the family Poaceae conducted elsewhere (Wang et al., 2015). However, the *U. mosambicensis* genome exhibited a higher CG content compared to the 13 species evaluated by Deng et al. (2016). This suggests that *U. mosambicensis* may present greater genome stability, likely due to the presence of triple hydrogen bonds, which could potentially influence the mutation rates of the genes associated with or containing microsatellites.

One of the main advantages of using microsatellites compared to other molecular markers is their wide distribution across various components of the genome, including coding sequences (CDS), introns, exons, and untranslated regions (UTRs). However, in plants of the Poaceae family, microsatellites are predominantly found in UTR regions, highlighting their significance in the regulation of gene expression (Wang et al., 2015; Liu et al., 2016). While recombination may influence the relative abundance of microsatellites, it remains unclear whether they are a cause or consequence of this phenomenon (Deng et al., 2016). Additionally, it is known that, besides slippage caused by DNA polymerase, uneven recombination plays a crucial role in the expansion and contraction of microsatellite repeats in the studied genomes (Ellengreen, 2004).

Repeat* motifs distribution *of pure and interrupted microsatellites

Pure microsatellites were dominant in the *Urochloa mosambicensis* genome, showing a relative abundance of 80.54 loci/Mb and a density of 1,387.88 bp/Mb (Table 1). The three most abundant basic repetitive units were tri-, penta-, and dinucleotides, with relative abundances of 31.52, 16.62, and 13.69 loci/Mb, and relative densities of 558.66, 263.22, and 247.39 bp/Mb, respectively.

The highest mean number of repeats was observed in mononucleotides (13.30 ± 1.72), dinucleotides (9.04 ± 5.54) and trinucleotides (5.91 ± 1.72). Overall, a strong negative linear correlation ($r = 0.93$, $p < 0.001$) was found between the number of repeats and the length of the repeat unit (ranging from 1 to 6 bp).

Putative markers, excluding mononucleotide motifs, were designed using the Primer3 engine included into the Krait software, based on the following parameters: (*i*) a maximum melting temperature difference of 1°C between primers; (*ii*) minimum, optimal, and maximum primer lengths of 18, 20 and 27 bp, respectively; (*iii*) GC content ranging from 30 % and 80 %; and (*iv*) target product sizes between 100 to 200 bp. Under these conditions, hexanucleotides (87.18%) exhibited the highest percentage of amplifiable sequences, followed by trinucleotides (86.96%), tetranucleotides (83.75%), pentanucleotides (82.36%), and dinucleotides (73.07%), demonstrating their potential as reproducible molecular markers.

The higher abundance of trinucleotide motifs in the *U. mosambicensis* genome observed in this study is consistent with findings in several plant species, including monocots (Varsheney et al., 2005; Gao et al., 2013). Microsatellite motifs are widely distributed throughout the plant genome, including in organelles such as mitochondria and chloroplasts. The literature reports a high frequency of these motifs in coding regions, specifically in exons, as well as in the 5'-UTR and 3'-UTR, where they play critical roles in regulating transcription, translation, and chromatin structure (Gao et al., 2013; Vieira et al., 2016). The presence of microsatellites in coding regions (CDS) can disrupt the reading frame of genes, potentially leading to gain or loss of function and affecting regulatory elements associated with transcription and translation. This may promote rapid protein evolution and increased adaptive flexibility. Notably, studies indicate that trinucleotide microsatellites exhibit virtually no difference in mutation patterns between coding and non-coding regions (Gao et al., 2013).

Table 1. Occurrence of different types of SSR motifs in the *Urochloa mosambicensis* genome[a].

Class	Type	No.	Average no. repeats	Total length (bp)	Average length (bp)	Relative abundance (Loci/Mb)	Density (bp/Mb)
Pure	Mono-	188	13.30 (1.72)	2501	13.3	5.68	75.58
	Di-	453	9.04 (5.54)	8186	18.07	13.69	247.39
	Tri-	1043	5.91 (1.72)	18486	17.72	31.52	558.66
	Tetra-	80	4.41 (0.81)	1412	17.65	2.42	42.67
	Penta-	550	3.17 (0.58)	8710	15.84	16.62	263.22
	Hexa-	351	3.15 (0.50)	6630	18.89	10.61	200.36
	Total	2665	5.99 (3.85)	45925	17.24	80.54	1387.88
Interrupted pure[a]		16	22.00 (9.47)	889	55.56	0.48	26.87
Pure compound		38	16.34 (10.08)	1807	47.55	1.15	54.61
Interrupted compound		60	17.27 (12.26)	2837	47.28	1.81	85.73

[a] Data generated from the MSDB software.

The most abundant pure mononucleotide motif was 'A/T', with a relative abundance of 4.53 loci/Mb, and a density of 60.98 bp/Mb (Table 2). This pattern is also observed in monocot and dicot plants (Sonah et al., 2011). Due to the presence of only two hydrogen bonds between the DNA strands, A/T-rich mononucleotides are more unstable than G/C-rich motifs, making them more prone to mutations that can disrupt critical cellular processes, such as mRNA processing and genetic silencing regulated by the 5'-UTR and 3'-UTR regions in eukaryotes (Kalia et al., 2011).

In the forage species *U. ruziziensis*, the most prevalent motif identified was AG/CT, with a frequency of 3,717 loci/Mb and a density of 69,447 bp/Mb, followed by AT/AT, which had 2,659 loci/Mb and 43,880 bp/Mb (Silva et al., 2013). Similar findings have been reported in Embryophyta (Tóth et al., 2000) and in certain monocots, such as *Brachypodium distachyon* and *Sorghum bicolor*, particularly within the coding sequence (CDS) regions of their genomes (Sonah et al., 2011). These findings are consistent with research on the Triticaceae family, which similarly reported a high frequency of these motifs in the 5'-UTR regions (Deng et al., 2016; Oliveira et al., 2006).

Among trinucleotides, AGG/CCT and AAG/CTT were the most abundant, with a relative abundance of 2.115 and 1.028 loci/Mb, and density of 37.26 and 16.59 bp/Mb, respectively (Table 2), which differs from what is observed in Embryophyta, in which the most abundant motif was AAG/CTT (Tóth et al., 2000). In *Brachypodium*, AGG/CTT appears as the second most abundant motif, with 15.8 loci/Mb, behind CCG/CGG, with 32.1 loci/Mb. Studies indicate that GC-rich trinucleotides are more frequently located in exons, whereas AT-rich trinucleotides tend to be more widely distributed across all genomic components (Vieira et al., 2016).

Tetra-, penta-, and hexanucleotide microsatellites were found to be the least abundant motifs, as also observed in other plant-based studies (Tóth et al., 2000; Gao et al., 2013). Notably, our analysis revealed a higher abundance of pentanucleotides in the *U. mosambicensis* genome compared to other plant species. The predominant tetranucleotide motif was AAAT/ATTT, exhibiting a frequency of 0.18 loci/Mb and a density of 3.86 bp/Mb, with an average of 5.33 repeats. Trinucleotide or longer oligonucleotide markers are generally preferred for genotyping due to their ease of size differentiation during gel electrophoresis, allowing accurate allele scoring, and their lower susceptibility to slippage during PCR, thus avoiding the formation of stutter bands. Tetranucleotide markers with 11 to 16 repeats are, therefore, recommended as optimal for population studies (Asch et al., 2010). In SSR loci, an increase in the number of repeats is generally associated with higher mutation rates, and loci with more than 16 repeats have a greater likelihood of accumulating interrupted motifs, potentially compromising the interpretation of results. Furthermore, a large number of repeats can result in PCR failures, such as increased allele dropout (Selkoe and Toonen, 2006).

Contrary to the findings reported by Tóth et al., (2000), who observed that pentanucleotides repeats are the least common motifs across a broad range of eukaryotic genomes, our study found that pentanucleotides were the third most frequent motif type in *U. mosambicensis.* This discrepancy might reflect species-specific genomic features, differences in the genome composition of grasses, or the selected filtering parameters and sequencing platform used in our sequencing approach.

The most common were ACCGC/GCGGT and AAAAT/ATTTT, with relative abundances of 0.66 and 0.33 loci/Mb, and relative densities of 9.97 and 5.29 bp/Mb, respectively. In the Triticaceae family, a group of flowering plants within the Poaceae (grass) family, these microsatellites are predominantly found in coding sequences (Deng et al., 2016). In Embryophyta, the AAAAT motif is notably widespread across different genomic regions, including intergenic regions, exons, and introns (Tóth et al., 2000), as also seen in *Aristotelia chilensis* (Bastías et al., 2016). Pentanucleotides, along with tetra- and hexanucleotides, exhibit a lower tendency to stutter formation during PCR amplification (Guichoux et al., 2011).

AAAAAC/GTTTTT and ATCCTC/GAGGAT motifs were the most common among hexanucleotides, with abundance of 0.393 and 0.332 loci/Mb and relative density of 7.072 and 5.984 bp/Mb, respectively. In Embryophyta, AAAAAC is the second most frequent, followed by AAAAAT, mainly in the introns and intergenic regions (Tóth et al., 2000). Similarly to trinucleotides, hexanucleotides are usually more abundant in exons, where they are controlled by stronger mutation pressure than in other regions (Vieira et al., 2016).

Overall, the three most abundant repeat motifs identified in the forage grass *U. mosambicensis* genome were A/T (4.53 loci/Mb), AG/CT (3.72 loci/Mb), and AT/AT (2.66 loci/Mb), as shown in Figure 1.Figure 2 illustrates an example of an imperfect microsatellite sequence used to evaluate the degree of substitutions, insertions, and deletions. The alignment, generated by the Krait algorithm to identify imperfect microsatellites within contigs, revealed that the AGTTT motif spans 110 nucleotides, with 100 matching bases. The sequence also exhibited 7 substitutions, 3 insertions, 4 deletions, and an identity score of 0.877.

Table 2. Contents of the first four motifs for each microsatellite type in the *Urochloa mosambicensis* genome.

Type	Motif	N	Length (bp)	Number of tandem repeats	Average length (bp)	Abundance (loci/Mb)	Density (bp/Mb)
Mono-	A/T	150	2018	13.450	13.450	4.530	60.985
	C/G	38	483	12.710	12.710	1.150	14.597
Di-	AC/GT	36	622	8.639	17.278	1.088	18.797
	AG/CT	123	2298	9.341	18.683	3.717	69.447
	AT/AT	88	1452	8.250	16.500	2.659	43.880
	CG/CG	9	108	6.000	12.000	0.272	3.264
Tri-	AAG/CTT	34	549	5.382	16.147	1.028	16.591
	ACC/GGT	13	207	5.308	15.923	0.393	6.256
	AGG/CCT	70	1233	5.871	17.614	2.115	37.262
	ATG/CAT	13	237	6.077	18.231	0.393	7.162
	CCG/CGG	29	522	6.000	18.000	0.876	15.775
Tetra-	ATAG/CTAT	1	16	4.000	16.000	0.030	0.484
	ACCG/CGGT	2	36	4.500	18.000	0.060	1.088
	AAAT/ATTT	6	128	5.333	21.333	0.181	3.868
	AAAC/GTTT	1	16	4.000	16.000	0.030	0.484
	AAAG/CTTT	3	48	4.000	16.000	0.091	1.451
Penta-	CCGCG/CGCGG	7	105	3.000	15.000	0.212	3.173
	ACGGC/GCCGT	5	95	3.800	19.000	0.151	2.871
	ACCGC/GCGGT	22	330	3.000	15.000	0.665	9.973
	AAAAG/CTTTT	10	155	3.100	15.500	0.302	4.684
	AAAAT/ATTTT	11	175	3.182	15.909	0.332	5.289
Hexa-	AAAAAC/GTTTTT	13	234	3.000	18.000	0.393	7.072
	AAAAAG/CTTTTT	4	78	3.250	19.500	0.121	2.357
	AACCCG/CGGGTT	4	72	3.000	18.000	0.121	2.176
	ATCCTC/GAGGAT	11	198	3.000	18.000	0.332	5.984
	ATCACC/GGTGAT	1	18	3.000	18.000	0.030	0.544

Imperfections in the *Urochloa mosambicensis* microsatellite loci were primarily caused by nucleotide substitutions (17.56%), followed by insertions (2.13%) and deletions (1.94%), as determined by the Krait algorithm (Figure 3). Mononucleotides (18.62%), dinucleotides (18.02%), and trinucleotides (17.11%) exhibited the highest rates of imperfections. In terms of substitutions, mononucleotides had the highest percentage at 18.62%, followed by dinucleotides (13.62%) and trinucleotides (13.49%). Hexanucleotides showed the highest rates of insertions, while dinucleotides had the highest frequency of deletions (Figure 3A).

Only 4% of the microsatellites were identified as imperfect, consisting of interrupted pure, compound, and interrupted compound types (Chambers and MacAvoy, 2000). Among these, interrupted pure microsatellites were the least abundant, comprising just 0.6%, with 0.48 loci/Mb and a relative density of 26.87 bp/Mb, as determined by the algorithm implemented in the MSDB software. The average number of tandem repeats was 22.00 ± 9.47 (Table 1). Interruptions were more frequent in di- and trinucleotides, occurring in 50% and 37.5% of cases, with average lengths of 17.54 bp and 15.66 bp, and abundances of 0.24 and 0.18, respectively (Figure 3B). No interruptions were observed in tetra- and hexanucleotide microsatellites. A single mononucleotide, $(C)_{13}GGGG(C)_{14}$, and a pentanucleotide, $(AACTA)_3ACT(AACTA)_3$, each had only one interruption. The highest number of tandem repeats was found in dinucleotides (e.g., $(CA)_8CG(CA)_{77}$), while the lowest was in pentanucleotides (e.g. $(AACTA)_3ACT(AACTA)_3$).

The structural analysis revealed the presence of 38 distinct pure compound microsatellite loci and 60 interrupted compound loci. These two classes of motifs exhibited relative abundances of 1.15 and 1.81 loci/Mb, relative densities of 64.61 and 85.73 bp/Mb, and average numbers of tandem repeats of 16.34 ± 10.08 and 17.27 ± 12.26, respectively. For the first class, the structures with the highest number of repeats were $(TC)_{29}(AC)_{25}$ and $(TA)_{20}(GA)_{15}$. For interrupted compounds, the highest number of repeats was found in $(CT)_7(TC)_{29}(AC)_{25}CC(CA)_{24}$. Although the origin of most compound microsatellites is not yet fully understood, it is suggested that they arise from duplicated imperfections during elongation, particularly at the ends of primary microsatellites, with a higher frequency of these microsatellites observed in coding regions compared to non-coding regions across various species (Kofler et al., 2008). Our results also corroborate these findings, showing high frequencies of structures composed of two adjacent microsatellites (diSSR) at 86.84%, followed by triSSR at 10.53% and tetraSSR at 2.63%. Structures with more than four adjacent microsatellites were not analyzed in this study.

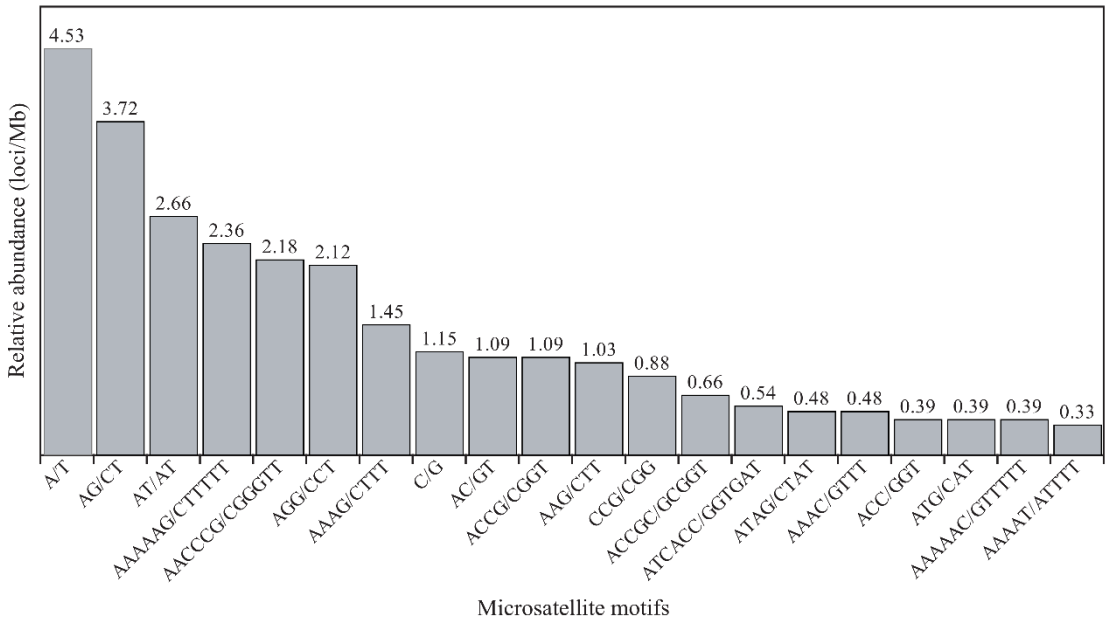

Fig 1. Relative abundance of the 20 most frequent pure microsatellite motifs in the *Urochloa mosambicensis* genome.

Although compound microsatellites are more commonly found in coding regions, the cSSRs identified in this study could serve as promising candidate molecular markers for plant population studies, as cSSRs have been proven to be highly polymorphic (Kumar et al., 2009; Zhai et al., 2010; Xu et al., 2011; Wang et al., 2014; Sen et al., 2017; Basu and Bhattacharya, 2022).

When analyzing the results generated by the MSDB program, we observed a trend, albeit not statistically significant, indicating that an increase in the number of non-consensus bases in interrupted pure microsatellites was associated with a reduction in the number of microsatellite repeats, showing a linear correlation of -0.447 ($p = 0.082$) (Figure 4A). This finding further supports the idea that the expansion of microsatellite size may influence mutation rates and potentially affect the degree of polymorphism desired in population studies (Guichoux et al., 2011).

Considering a maximum distance of 10 bp between adjacent microsatellites, a significant correlation was observed between the percentage of gaps ($r = -0.475$, $p = 0.001$) and the total number of repeats in the structure of interrupted compound microsatellites in *Urochloa mosambicensis*. This finding confirms that these interruptions can also hinder the expansion of microsatellite size (Figure 4B).

To determine whether microsatellites in *Urochloa mosambicensis* are randomly distributed or clustered within the genome, we analyzed the microsatellites separately, considering a dMax of 99 bp, the maximum number generated by the Krait program among diSSRs. In contrast to the findings of Kofler et al. (2008), the correlation between the frequency of compound structures and the dMax (gaps) between microsatellites was -0.15 ($p = 0.30$) (Figure 4C). Although this result is not statistically significant, it suggests a tendency for microsatellites to cluster within the genome of *U. mosambicensis*.

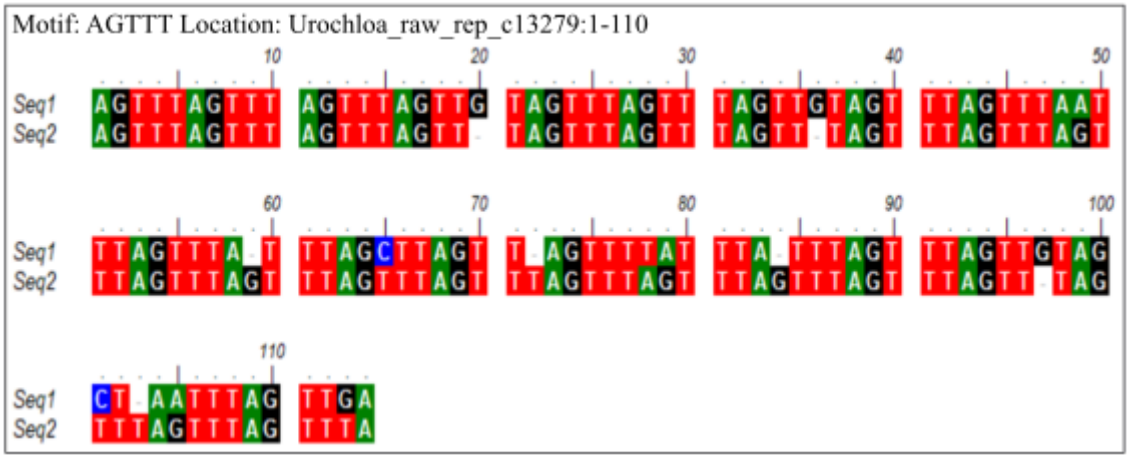

Fig 2. Example of an imperfect microsatellite sequence (colored strand), contig Urochloa_raw_rep_c132279:1-110.

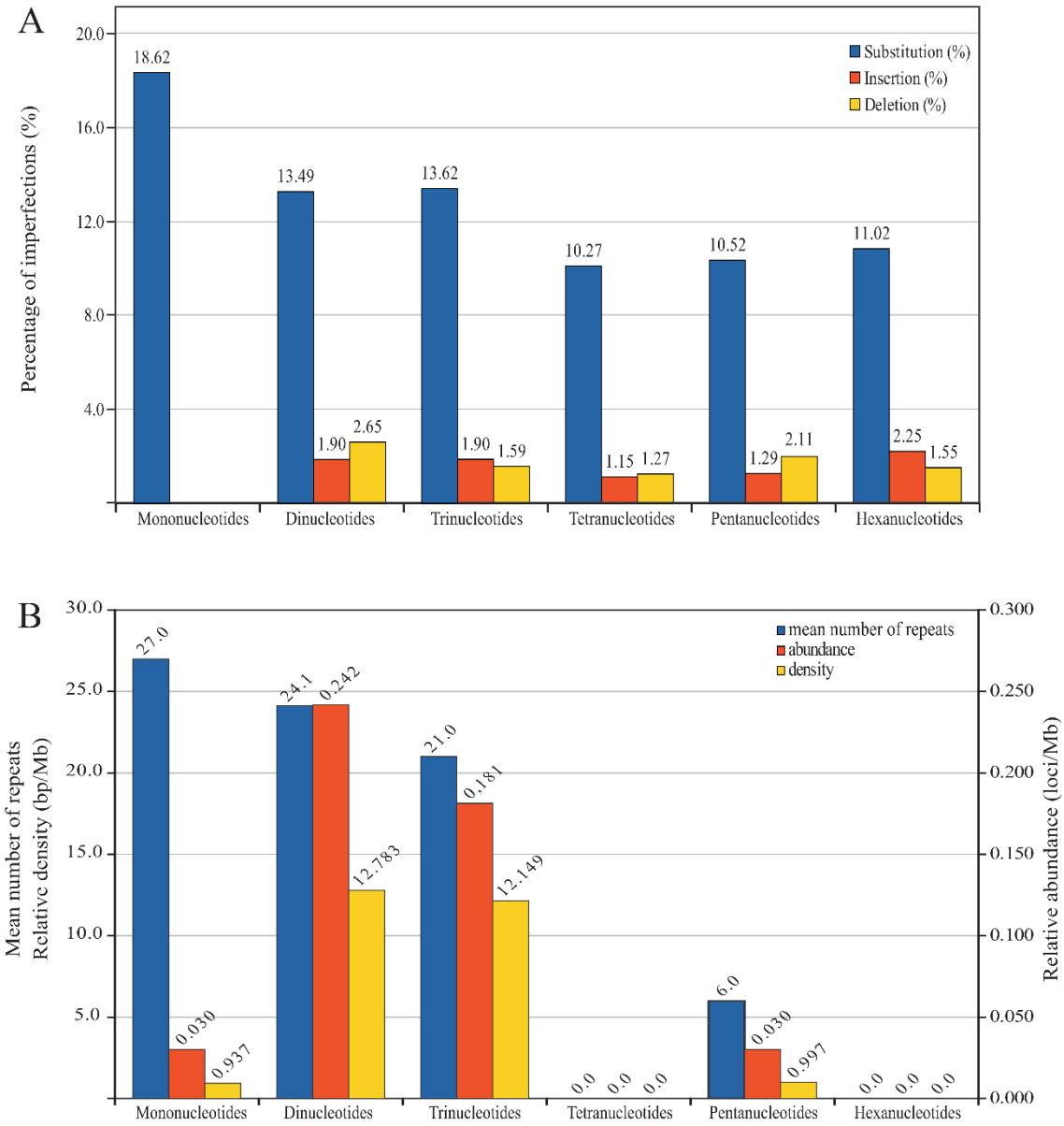

Fig 3. Distribution of imperfections (substitutions, insertions, and deletions of nucleotides) in the alignment of pure microsatellites across mono-, di-, tri-, tetra-, penta-, and hexanucleotides, analyzed using the Krait software (A). The distribution of interrupted pure microsatellites, along with the average number of repeats, abundance, and relative density, calculated using the MSDB program (B).

Materials and Methods

Plant materials and genomic DNA isolation

Urochloa mosambicensis genotypes were obtained from Embrapa Beef Cattle and maintained at Embrapa Goats and Sheep in Sobral, CE, Brazil (3º45'03.55" S, 40º20'37.45" W). Young leaves were collected, and genomic DNA was extracted using the ExtractME® Genomic DNA Kit (BLIRT DNA, Gdansk, Poland) according to the manufacturer's instructions. DNA quantity and quality were assessed on 0.8% agarose gels.

Library preparation and next generation sequencing

A single individual with high DNA yield and quality was selected for paired-end library DNA preparation, following the protocol provided in the Illumina Nextera XT Library Preparation Kit (Illumina, San Diego, CA, USA). Approximately 1 mg of genomic DNA (gDNA) was tagged and fragmented using the Nextera XT Transposome, followed by limited-cycle PCR amplification, AMPure XP magnetic-bead purification (Agencourt Bioscience Corporation, Beverly, MA, USA), and the Illumina Nextera XT bead-based normalization protocol. The DNA library was then sequenced on a MiSeq Benchtop Sequencer (Illumina Inc., San Diego, CA), targeting 500-bp fragments with 2 × 250-bp reads in a paired-end sequencing configuration.

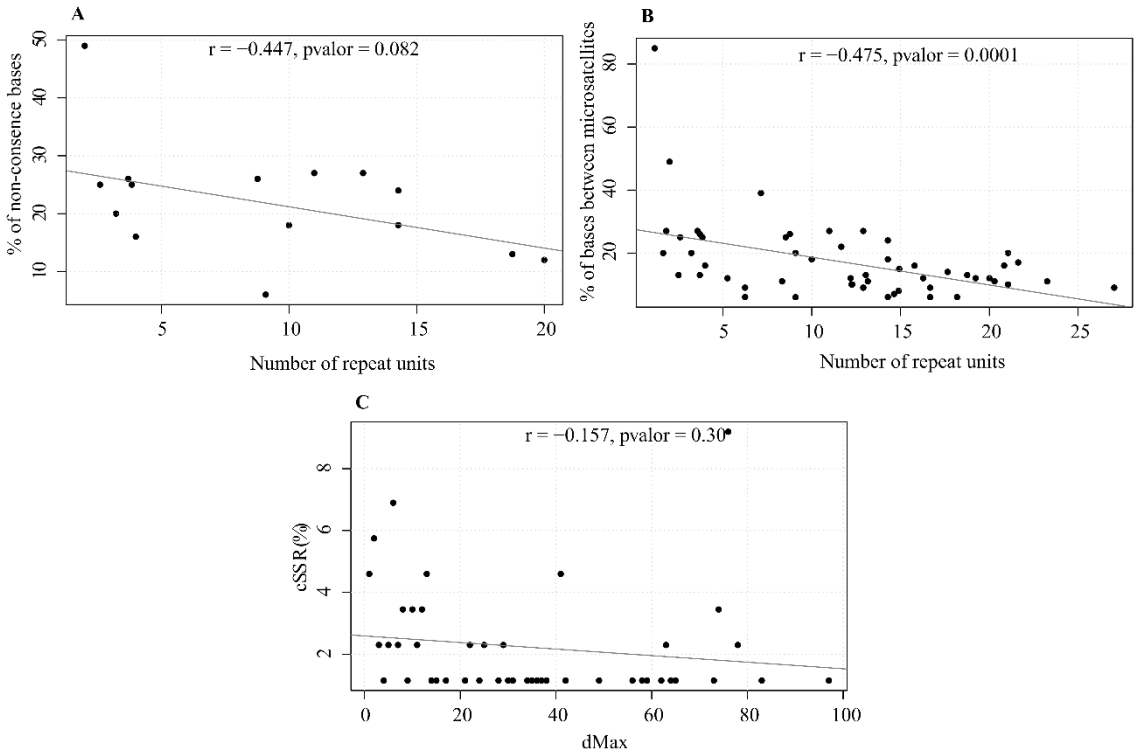

Fig 4. (A) Correlation between the percentage of non-consensus bases and the average number of microsatellite repeats in *Urochloa mosambicensis* for interrupted pure microsatellites, as analyzed using the MSDB program. (B) Correlation between the percentage of bases in the structures (i.e., dMax) that comprise interrupted compound microsatellites and the average number of tandem repeats, with a maximum distance limit of 10 bp. (C) Correlation between dMax and the percentage of cSSRs generated at each distance, as determined by the Krait program, with a maximum distance threshold set at 99 bp.

Sequencing data processing, genome assembly and SSR mining

Sequence data were pre-processed for quality assessment using the FastQC pipeline (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) prior to contig assembly, as the Illumina MiSeq platform may generate a small proportion of low-quality reads. Quality trimming was performed on the ends of low-quality sequences (sequencing quality value < 20). Additionally, any remaining adapters and reads containing more than 10% ambiguous bases (Ns) were removed. High-quality sequence reads were then assembled into contigs using CLC Genomics Workbench v.7.0.4 (Qiagen, Carlsbad, CA, USA).

Contigs containing simple sequence repeats were identified using the Krait software v1.0.3 (Du et al., 2018). SSR loci were classified into four categories: pure, interrupted pure, pure compound (distinct and adjacent SSRs), and interrupted compound motifs, with repeat units ranging from one to six nucleotides. The length criteria were set as follows: ≥ 12 nucleotides for mononucleotides, ≥ 6 for dinucleotides, ≥ 5 for trinucleotides, ≥ 4 for tetranucleotides, ≥ 3 repeats for penta- and hexanucleotides. These thresholds are consistent with previous studies (Mun et al., 2006) and ensure that the repeats are flanked by at least 100 bp on both sides, facilitating primer design.

To investigate interrupted compound SSRs, the maximum allowed distance between two adjacent microsatellites (dMax) was set to 10 bases. The MSDB v2.4.3 software (Du et al., 2013) was also used to analyse the statistical profile of interrupted pure microsatellites (e.g. AGAGAGctgAGAGAGAG).

Given that DNA is a double-stranded molecule and that the start site of an SSR can be arbitrarily chosen, complementary SSR sequences and circular permutations were grouped into the same class (Jurka and Pethiyagoda, 1995). For example, the sequences AGT, GTA, TAG, TCA, CAT, and ATC were treated as equivalent motifs in the analysis.

All SSR categories were analyzed based on for the number of motif type, length variation, and frequency of occurrence. The total length (bp) of identified SSR sequences, total SSR count, average motif length, relative abundance (number of SSR motifs per megabase, Mb), and relative density (total motif length in bp per Mb) were calculated to facilitate comparisons among different repeat types or categories. The types and frequencies of major repeats, including mononucleotide motifs, are listed on the Results section.

Conclusion

This study provides new insights into the distribution of microsatellites in the *Urochloa mosambicensis* genome, supporting their application as molecular markers in genetic breeding and plant population studies. Among the different repeat classes, uninterrupted SSRs are the most suitable candidates for genetic marker development, particularly, trinucleotide and tetranucleotide repeats, as they are less prone to forming stutter bands or experiencing allele dropouts during PCR, among other advantages. Additionally, imperfect microsatellites, including interrupted pure and interrupted compound SSRs, may serve as more conserved markers, with potential applications in targeting candidate genes, as well as in phylogenetic studies.

Conflict of Interests

The authors declare that there is no conflict of interests.

Acknowledgments

The authors would like to thank the Genomics and Bioinformatics Unit of the Drug Research and Development Center at the Federal University of Ceará for technical support in sequencing efforts. This study was funded by the Brazilian Agricultural Research Corporation (Embrapa) under project grant No. 20.23.01.009.00.00 (PMG-*Urochloa*) and by the Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP) through grant No. BP5-0197-00214.01.00/22. FMD also acknowledges FUNCAP for the Research Productivity Grant. We thank the reviewers for their thoughtful feedback and valuable suggestions. Field collection and DNA access were authorized by IBAMA/CGEN under permit No. A78EDC6.

References

Abdelkrim J, Robersten BC, Stanton JA, Gemmell NJ (2009) Fast, cost-effective development of species-specific microsatellite markers by genome sequencing. Biotechniques 46:185–191.

Asch B, Pinheiro R, Pereira R, Alves C, Pereira V, Pereira F, Gusmão L, Amorim A (2010) A framework for the development of STR genotyping in domestic animal species: characterization and population study of 12 canine X-chromosome loci. Electrophoresis 31:303–308.

Bastías A, Correa F, Rojas P, Almadas R, Muñoz C, Sagredo B (2016) Identification and characterization of microsatellite loci in Maqui (*Aritotelia chilensis* [Molina] Stunz) using Next-Generation. PLoS One 11(7):e0159825.

Basu S, Bhattacharya S (2022) Development and characterization of EST-SSR derived functional domain marker (FDM) in *Phaseolus vulgaris* (common bean). Plant Omics J 15(1):17–24.

Bindler G, Van der Hoeven R, Gunduz I, Plieske J, Ganal M, Rossi L, Gadani F, Donini P (2007) A microsatellite marker-based linkage map of tobacco. Theor Appl Genet 114:341–349.

Chambers GK, MacAvoy ES (2000) Microsatellites: consensus and controversy. Comp Biochem Physiol B 126(4):455–476.

Chat V, Ferguson R, Morales L, Kirchhoff T (2022) Ultra low-coverage whole-genome sequencing as an alternative to genotyping arrays in genome-wide association studies. Front Genet 12:790445.

Coates DB (1997) The recovery of Nixon Sabi grass pastures following severe drought. Trop Grassl 31:67–72.

Danin-Poleg Y, Reis N, Baudracco-Arnas S, Pitrat M, Staub JE, Oliver M, Arus P, Vicente CM, Katzir N (2000) Simple sequence repeats in *Cucumis*: mapping and map merging. Genome 43(6):963–974.

Deng P, Wang M, Feng K, Cui L, Tong W, Song W, Nie X (2016) Genome-wide characterization of microsatellites in Triticeae species: abundance, distribution, and evolution. Sci Rep 6:32224.

Du L, Zhang C, Liu Q, Zhang X, Yue B (2018) Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. Bioinformatics 34(4):681–683.

Du L, Li Y, Xiuyue L, Yue B (2013) MSDB: a user-friendly program for reporting distribution and building databases of microsatellites from genome sequences. J Hered 104(1):154–157.

Ellegreen H (2004) Microsatellites: simple sequence with complex evolution. Nat Rev Genet 5:435–445.

Ercisli S, Ipek A, Barut E (2011) SSR marker-based DNA fingerprinting and cultivar identification of olives (*Olea europaea*). Biochem Genet 49:555–561.

Ferreira RCU, Cançado LJ, Valle CB, Chiari L, Souza AP (2016) Microsatellite loci for *Urochloa decumbens* (Stapf) R.D. Webster and cross-amplification in other *Urochloa* species. BMC Res Notes 9:152.

Gao C, Ren X, Mason AS, Wang JLW, Xiao M, Fu D (2013) Revisiting an important component of plant genomes: microsatellites. Funct Plant Biol 1–17.

García-Martínez S, Andreani L, García-Gusano M, Geuna F, Ruiz JJ (2006) Evaluation of amplified fragment length polymorphism and simple sequence repeats for tomato germplasm fingerprinting: utility for grouping closely related traditional cultivars. Genome 49(6):648–656.

Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, Petit J (2011) Current trends in microsatellite genotyping. Mol Ecol Resour 11:591–611.
Jungmann L, Sousa ACB, Paiva J, Francisco PM, Vigna BBZ, do Valle CB, Zucchi MI, Sousa ACB (2009b) Isolation and characterization of microsatellite markers for *Brachiaria brizantha* (Hochst. ex A. Rich.) Stap. Conserv Genet 10:1873.
Jungmann L, Vigna BBZ, Paiva J, Sousa ACB, do Valle CB, Laborda PR, Zucchi MI, de Souza AP (2009a) Development of microsatellite markers for *Brachiaria humidicola* (Rendle) Schweick. Conserv Genet Resour 1:475–479.
Jurka J, Pethiyagoda C (1995) Simple repetitive DNA sequences from primates: compilation and analysis. J Mol Evol 40:120–126.
Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK (2011) Microsatellite markers: an overview of the recent progress in plants. Euphytica 177:309–334.
Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167.
Kofler R, Schlotterer C, Luschutzky E, Lelley T (2008) Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. BMC Genomics 9:612.
Kumar P, Gupta VK, Misra AK, Modi DR, Pandey BK (2009) Potential of molecular markers in plant biotechnology. Plant Omics J 9(2):141–162.
Leão US, Bueno LG, Negreiros AB, Silva GR, Maggioni R, Britto FB, Sarmento JLR, Galvani DB, Diniz FM (2023) Unveiling the demographic background and genetic diversity of *Urochloa mosambicensis* (Poaceae) through genome-wide identification of simple sequence repeats and molecular marker development. Conserv Genet Resour 15:135–143.
Liu F, Hu Z, Liu W, Li J, Wang W, Liang Z, Wang F, Sun X (2016) Distribution, function and evolution characterization of microsatellite in *Sargassum thunbergii* (Fucales Phaeophyta) transcriptome and their application in marker development. Sci Rep 6:18947.
Lou RN, Jacobs A, Wilder AP, Therkildsen NO (2021) A beginner's guide to low-coverage whole genome sequencing for population genomics. Mol Ecol 30(23):5966–5993.
Mcivor JG (1984) Leaf growth and senescence in *Urochloa mosambicensis* and *U. oligotricha* in a seasonally dry tropical environment. Aust J Agric Res 35:177–187.
Mun JH, Kim DJ, Choi HK, Gish J, Debellé F, Mudge J, Denny R, Endré G, Saurat O, Dudez AM, Kiss GB, Roe B, Young ND, Cook DR (2006) Distribution of microsatellites in the genome of *Medicago truncatula*: a resource of genetic markers that integrate genetic and physical maps. Genetics 172(4):2541–2555.
Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira ML (2006) Origin, evolution and genome distribution of microsatellites. Genet Mol Biol 29(2):294–307.
Ray S, Satya P (2014) Next generation sequencing technologies for next generation plant breeding. Front Plant Sci 5:367.
Santos JC, Barreto MA, Oliveira FA, Vigna BB, Souza AP (2015) Microsatellite markers for *Urochloa humidicola* (Poaceae) and their transferability to other *Urochloa* species. BMC Res Notes 8:83.
Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol Lett 9:615–629.
Sen S, Dehury B, Sahu J, Rathi S, Yadav RNS (2017) In silico mining and characterization of simple sequence repeats (SSRs) from *Euphorbia esula* expressed sequence tags (ESTs): a potential crop for biofuel. Plant Omics J 10:53–63.
Silva PI, Martins AM, Gouvea EG, Pessoa-Filho M, Ferreira M (2013) Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads. BMC Genomics 14:17.
Sonah H, Deshmukh RK, Sharma A, Singh VP, Gupta DK, Gacche RN, Rana JC, Singh NK, Sharma TR (2011) Genome-wide distribution and organization of microsatellites in plants: An insight into marker development in *Brachypodium*. PLoS One 6(6):e21298.
Sui C, Wei J, Chen S, Chen H, Yang C (2009) Development of genomic SSR and potential EST-SSR markers in *Bupleurum chinense* DC. Afr J Biotechnol 8:8.
Taheri S, Lee Abdullah T, Yusop MR, Hanafi MM, Sahebi M, Azizi P, Shamshiri RR (2018) Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. Molecules 23:399.
Tautz D, Renz M (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res 12:4127–4138.
Tóth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10:967–981.
Treydte AC, Baumgartner S, Heitkönig IMA, Grant CC, Getz WM (2013) Herbaceous forage and selection patterns by ungulates across varying herbivore assemblages in a South African savanna. PLoS One 8(12):e82831.
Varsheney RK, Graner A, Sorrelis ME (2005) Genic microsatellite markers in plants: features and applications. Trends Biotechnol 23:1–9.
Vieira MLC, Santini L, Diniz AL, Munhoz CF (2016) Microsatellite markers: What they mean and why they are so useful. Genet Mol Biol 39:312–328.
Vigna BBZ, Alleoni GC, Jungmann L, do Valle CB, Souza AP (2011) New microsatellite markers developed from *Urochloa humidicola* (Poaceae) and cross amplification in different *Urochloa* species. BMC Res Notes 4:523.
Wang Y, Yang C, Jin Q, Zhou D, Wang S, Yu Y, Yang L (2015) Genome-wide distribution comparative and composition analysis of the SSRs in Poaceae. BMC Genet 16:18.
Wang Y-H, Chen N, Zhang Y-C, Fu C-X (2014) Development and characterization of microsatellite markers for the Chinese endangered medicinal plant *Tetrastigma hemsleyanum*. Genet Mol Res 13:9062–9067.

Weber JL (1990) Informativeness of human (dC-dA)n, (dG-dT)n polymorphisms. Genomics 7:524–530.
Xu X-H, Wan Y, Qi Z-C, Qiu Y-X, Fu C-X (2011) Isolation of compound microsatellite markers for the Mediterranean shrub *Smilax aspera* (Smilacaceae). Am J Bot e64–e66.
Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am J Bot 99:193–208.
Zhai S-N, Yan X-L, Nakamura K, Mishima M, Qiu Y-X (2010) Isolation of compound microsatellite markers for the endangered plant. *Neolitsea sericea* (Lauraceae). Am J Bot e139–e141.
Zhang Y, Yuan X, Teng W, Chen C, Wu J (2016) Identification and phylogenetic classification of *Pennisetum* (Poaceae) ornamental grasses based on SSR locus polymorphisms. Plant Mol Biol Rep 34:1181–1192.