Mammal Research (2020) 65:359–373
https://doi.org/10.1007/s13364-019-00458-x

ORIGINAL PAPER

Genome-wide characterization and analysis of microsatellite sequences in camelid species

Manee M. Manee[1,2,3] · Abdulmalek T. Algarni[1] · Sultan N. Alharbi[4] · Badr M. Al-Shomrani[1] · Mohanad A. Ibrahim[1] · Sarah A. Binghadir[1] · Mohamed B. Al-Fageeh[1]

Received: 6 June 2019 / Accepted: 21 September 2019 / Published online: 14 November 2019

Abstract

Microsatellites or simple sequence repeats (SSRs) are among the genetic markers most widely utilized in research. This includes applications in numerous fields such as genetic conservation, paternity testing, and molecular breeding. Though ordered draft genome assemblies of camels have been announced, including for the Arabian camel, systemic analysis of camel SSRs is still limited. The identification and development of informative and robust molecular SSR markers are essential for marker assisted breeding programs and paternity testing. Here we searched and compared perfect SSRs with 1–6 bp nucleotide motifs to characterize microsatellites for draft genome sequences of the Camelidae. We analyzed and compared the occurrence, relative abundance, relative density, and guanine-cytosine (GC) content in four taxonomically different camelid species: *Camelus dromedarius*, *C. bactrianus*, *C. ferus*, and *Vicugna pacos*. A total of 546762, 544494, 547974, and 437815 SSRs were mined, respectively. Mononucleotide SSRs were the most frequent in the four genomes, followed in descending order by di-, tetra-, tri-, penta-, and hexanucleotide SSRs. GC content was highest in dinucleotide SSRs and lowest in mononucleotide SSRs. Our results provide further evidence that SSRs are more abundant in noncoding regions than in coding regions. Similar distributions of microsatellites were found in all four species, which indicates that the pattern of microsatellites is conserved in family Camelidae.

Keywords Camel · Genome · Microsatellite · SSR abundance · Molecular marker

Introduction

Camelus dromedarius, often referred to as the Arabian camel, is one of the most important members of the family Camelidae. The dromedary is a heat stress-resistant animal (Manee et al. 2017) able to live in extreme harsh environments such as those of the Arabian Peninsula, and its adaptations to arid conditions are remarkable. For instance, camels are able to vary their body temperature from 34 to 41.7 °C, and can conserve water by not sweating (Al-Swailem et al. 2010). Additional members of the Camelidae include the Bactrian camel (*C. bactrianus*) in Asia and the llama (*Lama glama*) and alpaca (*Vicugna pacos*) in South America (Groeneveld et al. 2010; Wu et al. 2014), which play crucial roles in transportation and the provision of important products such as milk and meat. Given the economic value of camelid species, their genetic characterization is essential; in particular, implementing proper strategies for conserving animal genetic resources requires the evaluation of genetic diversity both within and among populations. Consequently, assessment of camel genetic diversity is important to help the development of breeding programs, which will facilitate improvements to camel productivity and identify genetically unique structures, furthering the ongoing conservation and utilization of these valuable animals.

Communicated by: Joanna Stojak

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13364-019-00458-x) contains supplementary material, which is available to authorized users.

✉ Mohamed B. Al-Fageeh
mfageeh@kacst.edu.sa

1 National Center for Biotechnology, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

2 Center of Excellence for Genomics, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

3 Institute of Bioinformatics, University of Georgia, Athens, GA, USA

4 National Center for Stem Cell Technology, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

As morphological traits are highly affected by environmental factors (Shehzad et al. 2009; Jugran et al. 2013; Last et al. 2014), morphological variation is not necessarily an accurate marker for genetic variation. Molecular markers are key resources for genetic investigations, as they complement morphological information and are informative at any developmental stage (Backes et al. 2003). Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs), are composed of short repetitive DNA sequences, 1–6 base pairs (bp) in length, and are widely distributed in many eukaryotic (Xu et al. 2016; Qi et al. 2015) and prokaryotic (Gur-Arie et al. 2000; Yang et al. 2003) genomes. Microsatellites undergo rapid contractions and expansions in different populations of the same species because of replication slippage (Huntley and Golding 2006), and thus are very useful markers for evaluating genetic diversity and DNA fingerprinting.

Variation in SSR lengths may also lead to changes in the local structure of DNA or protein sequences (Mrazek et al. 2007). Evidence shows that SSRs are distributed nonrandomly in genomes. Comparative analysis of *Arabidopsis thaliana* and *Oryza sativa* revealed that SSR distributions were nonrandomly distributed in different genomic regions, and varied widely in different gene regions (Lawson and Zhang 2006). SSRs are found in both coding and noncoding regions (Katti et al. 2001). However, SSRs are more abundant in noncoding regions than in exons (Hancock 1995), with trinucleotide and hexanucleotide SSRs being more abundant in coding regions (Borstnik 2002; Subramanian et al. 2003). Previous studies suggested that SSRs in promoter regions may affect gene expression, and SSRs in introns may influence gene transcription or mRNA splicing (Li et al. 2004).

The availability of draft whole genome sequences for several camel species provides the opportunity to perform post-genomic analysis to compare and assess the distribution of microsatellites across camel genomes (Bactrian Camels Genome Sequencing and Analysis Consortium et al. 2012; Wu et al. 2014). To the best of our knowledge, genome-wide characterization and analysis of perfect microsatellites in camels have not yet been reported. To date, there are four camelid species with draft genome sequences: *C. dromedarius*, *C. bactrianus*, *C. ferus*, and *Vicugna pacos*. This study aimed to screen the whole genomes of these four species for microsatellite identification. In particular, we detected and characterized SSRs and their motifs, and examined their distribution and variations in different genomic regions, which will facilitate studying the structure of the camel genome. This study will serve as a foundation for further research to develop camel-specific SSR markers.

Materials and methods

Data source

At the time of this study, only four camelid species (*C. dromedarius*, *C. bactrianus*, *C. ferus*, and *V. pacos*) were known to have draft genome sequences, which according to the genomic resources of the National Center of Biotechnology Information (NCBI) have been assembled at scaffold level. These four assemblies were used for the analysis of SSR distributions at the genomic level. Genome sequences in FASTA format and annotation information in GFF format were downloaded from the NCBI RefSeq database (Pruitt et al. 2012) through the Genomes FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/). The accession numbers were GCF_000767585.1 (NCBI Eukaryotic Genome Annotation Pipeline Release version 100), GCF_000767855.1 (100), GCF_000311805.1 (101) and GCF_000164845.2 (101), respectively.

Identification of microsatellites

The software `PERF` v0.2.5 (Avvaru et al. 2017) was utilized for genome-wide SSR mining. This tool is implemented in the Python programming language for detection of microsatellites from DNA sequences. However, camelid species have very large genomes (> 2 Gb). For this reason, the criteria utilized in this study to search for perfect SSRs were as follows: motif size of 1 to 6 nucleotides long using (*-m* option) and (*-M* option), and minimum repeat numbers restricted to 12 repeats for mononucleotides, seven repeats for dinucleotides, five repeats for trinucleotides, and four repeats for tetra-, penta-, and hexanucleotides, which were consistent with previous studies (Qi et al. 2015; Liu et al. 2017; Qi et al. 2018). All other settings were set as default. In this study, repeats with unit patterns being circular permutations and/or reverse complements were deemed as one type for statistical analysis (Jurka and Pethiyagoda 1995; Li et al. 2009a). For instance, the unit AGG denotes AGG, GAG, GGA, CCT, TCC, and CTC in different reading frames or on the complementary strand. Relative frequency and relative density were used to help conduct comparisons between different repeat types or motifs. Relative frequency is the number of SSRs per megabase pair (Mb) of target sequence, and relative density is the length of SSRs in base pairs (bp) per Mb of the target sequence (Karaoglu et al. 2005). Total numbers of SSRs were normalized as relative frequency and relative density to perform comparisons between microsatellite sequences of different sizes.

Assigning microsatellites to genomic compartments

The sequences and coordinates of gene models, exons, coding sequences (CDSs), and intronic and intergenic regions for the four camelid genomes were determined according to the positions in the genome annotation files in GFF format downloaded from the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/). These GFF files were converted to BED files for further analysis using gff2bed (v2.4.28) (Neph et al. 2012). The draft genome sequences in FASTA format were indexed using the samtools faidx function implemented in SAMtools v1.7 (Li et al. 2009b). Intergenic and intronic coordinates were obtained using BEDtools subtract tool v2.26.0 (Quinlan and Hall 2010). Intergenic regions were defined as the interval sequences between genes, and intronic regions were defined as the interval sequences between exonic regions. Identified microsatellites were assigned to genomic compartments using the BEDtools intersect tool v2.26.0 (Quinlan and Hall 2010). Each tool was run with default settings.

Statistical analysis

All graphical and statistical analyses were conducted in the R programming environment (version 3.4.3) (R Core Team, 2017). The cor.test method='pearson' was used to elucidate correlations between SSR data sets, including relative frequency, relative density, and GC content.

Results

Identification and characterization of microsatellites in camelid genomes

We analyzed perfect SSRs from four draft camelid genomes (*C. dromedarius*, *C. bactrianus*, *C. ferus*, and *V. pacos*). Genome characteristics including genome size, GC content, number of SSRs, relative frequency, and relative density are summarized in Table 1. Perfect microsatellites were searched for and analyzed using PERF software. In total, 546762, 544494, 547974, and 437815 perfect SSRs were identified per genome, with overall frequencies of ~ 273 SSRs/Mb in *Camelus* genomes and 201.55 SSRs/Mb in *V. pacos*, accounting for approximately 0.52% and 0.37% of the genomes, respectively. The number of SSRs was positively correlated with relative frequency (Pearson, $r = 0.999$, $P < 0.01$) and GC content of SSRs across species (Pearson, $r = 0.979$, $P < 0.05$), but negatively correlated with genome size (Pearson $r = -0.994$, $P < 0.01$). Relative frequency and relative density of SSRs were also negatively correlated with genome size (Pearson, $r = -0.997$, $P < 0.01$ and Pearson, $r = -0.971$, $P < 0.05$, respectively). For instance, *V. pacos* has the largest genome (2172.21 Mb) among those surveyed, and was found to have the lowest SSR frequency and density (201.55 SSRs/Mb and 3828.30 bp/Mb, respectively).

The number, relative frequency, and density of perfect mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide repeat types for the four genomes are shown in Table 2. The results revealed that the relative frequencies and densities of a given type of microsatellites are greatly similar in these species (Fig. 1b, c), with the exception of the relative frequency and density of mononucleotide SSRs in *V. pacos*. The proportions of mono- to hexanucleotide SSRs were similar across the four genomes, particularly between *C. dromedarius*, *C. bactrianus*, and *C. ferus* (Fig. 1a). Mononucleotide SSRs were the most frequent type, followed by di-, tetra-, tri-, penta-, and hexanucleotide SSRs in decreasing order. Mononucleotide SSRs had frequencies of 69.16–135.79 SSRs/Mb and the highest densities of 951.09–2066.54 bp/Mb, accounting for 34.31–49.79% of the total number of SSRs. Hexanucleotide SSRs were the least frequent, only accounting for 0.76–1.00% of all SSRs.

Table 1 Overview of the four camelid genomes

Parameter	*C. dromedarius*	*C. bactrianus*	*C. ferus*	*V. pacos*
Genome size (Mb)	2004.06	1992.66	2009.19	2172.21
GC content (%)	40.82	41.04	40.79	39.65
Number of SSRs	546762	544494	547974	437815
Total length of SSRs (bp)	10551766	10109025	10742267	8315872
Frequency (SSRs/Mb)	272.83	273.25	272.73	201.55
Density (bp/Mb)	5265.18	5073.12	5346.55	3828.30
Genome SSRs content (%)	0.53	0.51	0.53	0.38

Table 2 Number, length, frequency, and density of mono- to hexanucleotide repeats in four camelid genomes

Repeat type	Parameter	*C. dromedarius*	*C. bactrianus*	*C. ferus*	*V. pacos*
Mono-	Number of SSRs	272044	269115	272822	150228
	Total length (bp)	4121293	4082129	4152080	2065967
	Frequency (SSRs/Mb)	135.75	135.05	135.79	69.16
	Density (bp/Mb)	2056.47	2048.58	2066.54	951.09
Di-	Number of SSRs	139273	136582	138305	148597
	Total length (bp)	3016070	2952242	3070860	3227802
	Frequency (SSRs/Mb)	69.50	68.54	68.84	68.41
	Density (bp/Mb)	1504.98	1481.56	1528.40	1485.95
Tri-	Number of SSRs	30536	30632	30565	31726
	Total length (bp)	687393	616659	746760	628686
	Frequency (SSRs/Mb)	15.24	15.37	15.21	14.61
	Density (bp/Mb)	343.00	309.47	371.67	289.42
Tetra-	Number of SSRs	86685	89197	87570	88526
	Total length (bp)	2194620	1964168	2183340	1920420
	Frequency (SSRs/Mb)	43.25	44.76	43.58	40.75
	Density (bp/Mb)	1095.09	985.70	1086.67	884.09
Penta-	Number of SSRs	14090	14546	14355	14378
	Total length (bp)	402550	367485	443175	349145
	Frequency (SSRs/Mb)	7.03	7.30	7.14	6.62
	Density (bp/Mb)	200.87	184.42	220.57	160.73
Hexa-	Number of SSRs	4134	4422	4357	4360
	Total length (bp)	129840	126342	146052	123852
	Frequency (SSRs/Mb)	2.06	2.22	2.17	2.01
	Density (bp/Mb)	64.79	63.40	72.69	57.02

GC content and adenine-thymine (AT) content were investigated in camelid SSRs. The overall GC contents of SSRs were almost identical for *C. dromedarius*, *C. bactrianus*, and *C. ferus*, accounting for approximately 22%, and slightly higher in *V. pacos* (∼ 26%). The lengths and proportions of GC and AT content of all SSR types are presented in Table 3 and Fig. 1d. From the results, we can observe that all SSR repeat types had high AT contents. Mononucleotide SSRs had the highest AT content (> 94%), followed in decreasing order by penta-, tetra-,

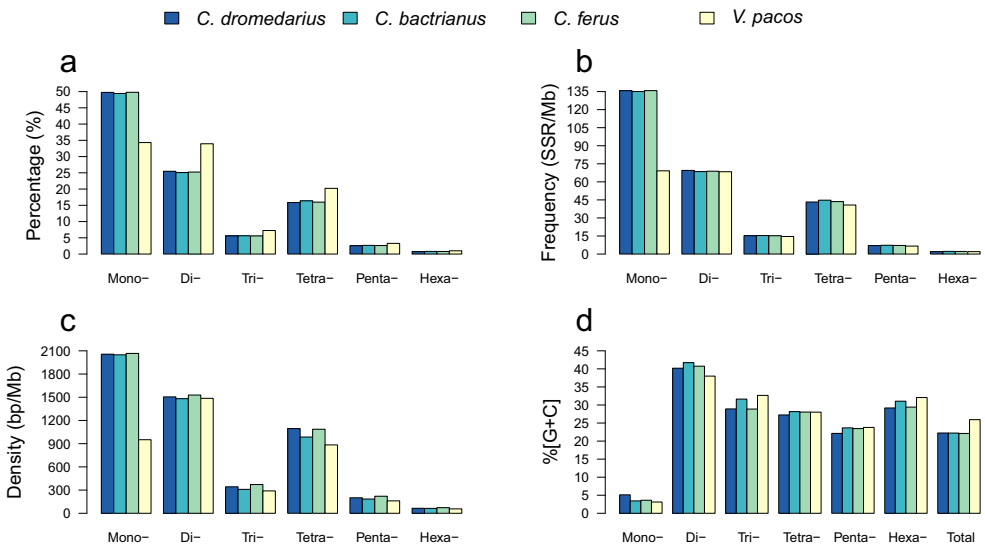

Fig. 1 Comparison of percentage, frequency, density, and GC content of SSRs in the camelid genomes. Percentages were calculated according to the total number of each SSR type divided by the total number of SSRs for that species. ABCD represent percentage, frequency, density, and GC content of SSRs, respectively

hexa-, trinucleotide, and the least being dinucleotide SSRs. The highest GC content among SSR repeat types was in the dinucleotide SSRs (~ 40%), and the least was in the mononucleotide SSRs (~ 4%) (Fig. 1d). The GC contents in tri- and hexanucleotide SSRs were highly similar across the four genomes, ranging from ~ 28 to ~ 32%. Interestingly, GC content in all SSR repeat types was significantly lower than that of the entire genome, except in dinucleotide SSRs. Furthermore, we conducted additional analyses to report all perfect SSRs in the four camelid genomes without applying any search criteria (supplementary files S1–S4).

Repeat numbers for different microsatellite types

The number of repeats in each SSR and the maximum repeats of each SSR type were found to be highly diverse in different microsatellite types across the four genomes. In general, the corresponding repeat motifs were almost identical between the four genomes, with the exception of fewer repeats for mononucleotide SSRs in *V. pacos* (Fig. 2).

Diversity of microsatellite motifs in camelid genomes

As noted above, the SSRs in camelid genomes were relatively AT-rich. To better understand why this is, we analyzed the motif composition of camelid SSRs. The most frequent SSR motifs for each repeat length were found to vary at the whole genome level across the four camelid species (Table 4). The major repeat motif types shared by the four genomes and having over 5000 SSRs were $(A)_n$, $(C)_n$, $(AC)_n$, $(AT)_n$, $(AG)_n$, $(AAT)_n$, $(AAC)_n$, $(AAAT)_n$, $(AAAC)_n$, $(AAAG)_n$, $(AAGG)_n$, $(AATG)_n$, $(AGAT)_n$, and $(AAAAC)_n$. The numbers of degenerate repeat motifs were found to be 2, 4, 10, and 33 for *C. dromedarius*, *C. bactrianus*, *C. ferus*, and *V. pacos*, respectively, and were identical between the four camelid genomes for mono- to tetranucleotide repeat types but different for pentanucleotide and hexanucleotide repeat types.

The predominant mononucleotide motif was $(A)_n$, accounting for 95–97% of the total mononucleotide SSRs in each genome (Fig. 3a). The $(C)_n$ repeat was the least frequent, with frequencies of less than 7 SSRs/Mb. In particular, *V. pacos* had approximately two-fold and one-fold lower frequency of $(C)_n$ repeats than *C. dromedarius*, *C. bactrianus*, and *C. ferus* (Table 4). The $(AC)_n$ repeat motif was the predominant dinucleotide SSR, occupying ~ 60% of all dinucleotide SSRs in the four genomes (Fig. 3b). The $(AT)_n$ repeat was the second most frequent dinucleotide repeat, with frequencies of 14.70–17.72 SSRs/Mb. The $(AG)_n$ motif was less abundant than $(AT)_n$, and $(CG)_n$ was the least frequent dinucleotide SSR. $(AAT)_n$ and $(AAC)_n$ motifs were the most frequent trinucleotide SSRs, together accounting for 49–53% of trinucleotide SSRs in the four camelid genomes (Fig. 3c). The third most frequent repeat motif was $(AGG)_n$, followed by $(ATC)_n$ and $(ACC)_n$, which had almost identical frequencies of approximately 1.50 SSRs/Mb. The $(ACG)_n$ motif was the least abundant trinucleotide SSR in the four camelid genomes.

Among tetranucleotide repeats, $(AAAT)_n$ and $(AAAC)_n$ were the most abundant with almost identical frequencies of approximately 8 SSRs/Mb, together accounting for 38.09–39.51% of total tetranucleotide SSRs in the four genomes (Fig. 3d). The third most frequent tetranucleotide motif was $(AAAG)_n$, with a similar frequency of more than 5 SSRs/Mb in these genomes, followed by the $(AAGG)_n$, $(AATG)_n$, and $(AGAT)_n$ motifs with frequencies ranging

Table 3 AT and GC content of SSRs for each SSR type in four camelid genomes

		C. dromedarius		*C. bactrianus*		*C. ferus*		*V. pacos*	
Type	Parameter	Length (bp)	%	Length (bp)	%	Length (bp)	%	Length (bp)	%
Mono-	A + T	3910440	94.88	3942002	96.57	4001972	96.38	2001376	96.87
	G + C	210853	5.12	140127	3.43	150108	3.62	64591	3.13
Di-	A + T	1803963	59.81	1720605	58.28	1819528	59.25	2001320	62.00
	G + C	1212107	40.19	1231637	41.72	1251332	40.75	1226482	38.00
Tri-	A + T	488647	71.09	421598	68.37	531143	71.13	423393	67.35
	G + C	198746	28.91	195061	31.63	215617	28.87	205293	32.65
Tetra-	A + T	1596792	72.76	1410982	71.84	1571157	71.96	1382423	71.99
	G + C	597828	27.24	553186	28.16	612183	28.04	537997	28.01
Penta-	A + T	313461	77.87	280545	76.34	339210	76.54	266107	76.22
	G + C	89089	22.13	86940	23.66	103965	23.46	83038	23.78
Hexa-	A + T	91965	70.83	87115	68.95	103099	70.59	84143	67.94
	G + C	37875	29.17	39227	31.05	42953	29.41	39709	32.06
Total	A + T	8205268	77.76	7862847	77.78	8366109	77.88	6158762	74.06
	G + C	2346498	22.24	2246178	22.22	2376158	22.12	2157110	25.94

from 2.47 to 4.28 SSRs/Mb. For pentanucleotide repeats, $(AAAAC)_n$ was the most abundant motif, occupying 44.30–47.17% of pentanucleotide SSRs in the camelid genomes (Fig. 3e). The second most frequent pentanucleotide motif was $(AAAAT)_n$, followed by $(AAAAG)_n$; these had almost identical frequencies of approximately 1 SSR/Mb, and together accounted for 28.09–28.83% of pentanucleotide SSRs in the four genomes. Hexanucleotide repeats were found to have a lower frequency and density compared to other microsatellite types. The predominant hexanucleotide motif was $(AAAAAC)_n$, with frequencies below 0.84 SSRs/Mb and densities below 24.06 bp/Mb, accounting for ~ 37% of hexanucleotide SSRs in *Camelus* species and 32.09% in *V. pacos*, followed by the $(AAAAAG)_n$ and $(AGATAT)_n$ motifs (Fig. 3f).

Distribution and motif diversity of microsatellites in different genomic regions

A microsatellite search was carried out in exons, CDSs, and intronic and intergenic regions to determine the distribution of SSRs in different genomic regions of *C. dromedarius*, *C. bactrianus*, *C. ferus*, and *V. pacos*. The comparison results revealed high similarity by region across the four genomes in terms of the relative abundances, densities, and percentages of most of the similar mono- to hexanucleotide SSRs; however, the occurrences and relative frequencies and densities of SSRs were found to differ significantly in coding and noncoding regions (Fig. 4). SSRs were most commonly located in intergenic regions, followed in order by intronic regions, exons, and CDSs (Fig. 4b). The frequencies of SSRs in CDSs of the four camelid species ranged from 0.83 to 1.26 SSRs/Mb, accounting for 0.30–0.36% of SSRs in *Camelus* species and 0.62% in *V. pacos*. The frequencies in exons ranged from 2.79 to 3.93 SSRs/Mb, accounting for 1.01, 1.28, 1.42, and 1.74% of SSRs in *C. dromedarius*, *C. bactrianus*, *C. ferus*, and *V. pacos*, respectively (Fig. 4a, b). The frequencies of SSRs in intergenic regions were 172.06, 170.45, 173.72, and 130.02 SSRs/Mb, respectively, accounting for ~ 62% of SSRs in all four species, while the frequencies in intronic regions were 99.69, 101.46, 97.90, and 70.37 SSRs/Mb, accounting for ~ 35% of SSRs in all four species (Fig. 4a, b). The respective densities of SSRs in coding regions were 14.93, 17.73, 20.14, and 24.15 bp/Mb for CDSs and 49.04, 60.99, 70.65, and 63.01 bp/Mb for exons (Fig. 4c). The densities of SSRs in noncoding regions were much higher, with intronic regions having densities of 1878.09, 1856.92,

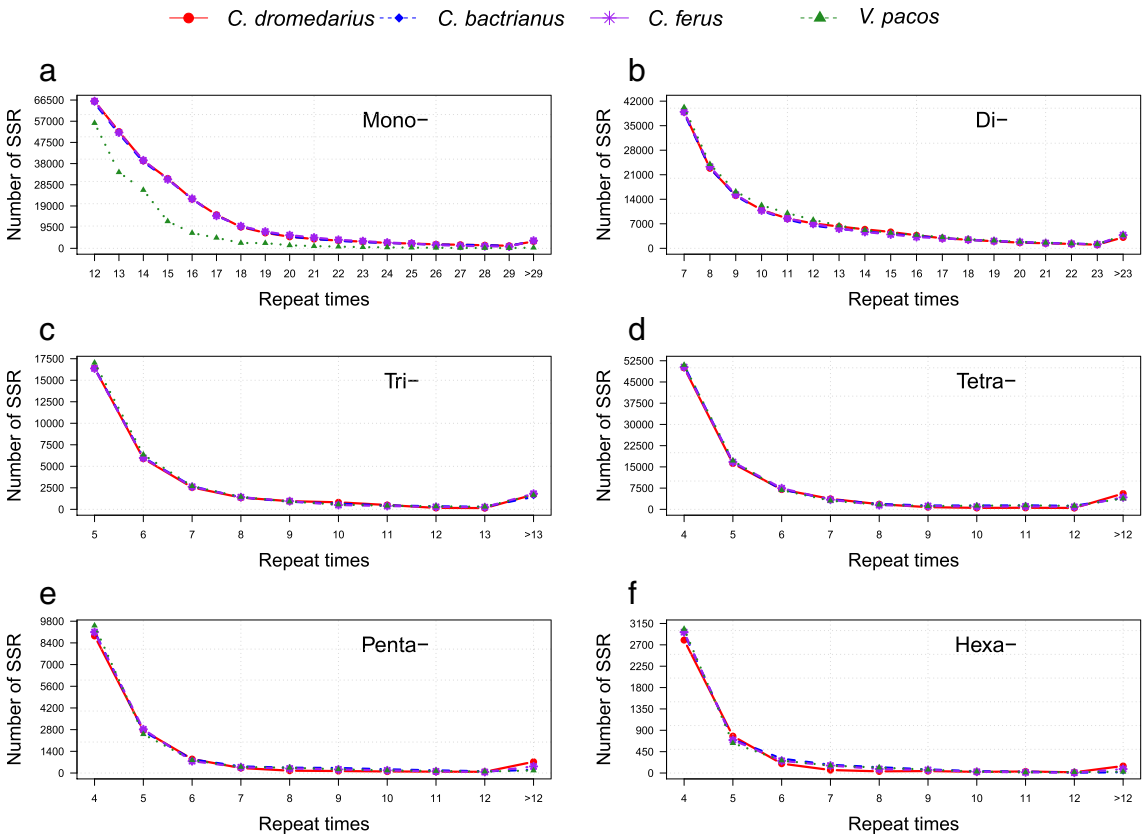

Fig. 2 Repeat times of different SSR types in the camelid genomes. ABCDEF represent mono-, di-, tri-, tetra-, penta-, and hexanucleotide SSR types, respectively

Table 4 The number, length, frequency, and density of the most frequent motifs for each SSR type in four camelid genomes

Repeat motif type	Parameter	*C. dromedarius*	*C. bactrianus*	*C. ferus*	*V. pacos*
A	Number of SSRs	258597	259391	263148	145207
	Total length (bp)	3910440	3942002	4001972	2001376
	Frequency (SSRs/Mb)	129.04	130.17	130.97	66.85
	Density (bp/Mb)	1951.26	1978.26	1991.83	921.36
C	Number of SSRs	13447	9724	9674	5021
	Total length (bp)	210853	140127	150108	64591
	Frequency (SSRs/Mb)	6.71	4.88	4.81	2.31
	Density (bp/Mb)	105.21	70.32	74.71	29.74
AC	Number of SSRs	86893	87077	88351	89060
	Total length (bp)	2039566	2075360	2099126	2055348
	Frequency (SSRs/Mb)	43.36	43.70	43.97	41.00
	Density (bp/Mb)	1017.72	1041.50	1044.76	946.20
AT	Number of SSRs	32512	29297	29691	38490
	Total length (bp)	598748	497606	575574	783302
	Frequency (SSRs/Mb)	16.22	14.70	14.78	17.72
	Density (bp/Mb)	298.77	249.72	286.47	360.60
AG	Number of SSRs	19424	19663	19789	20530
	Total length (bp)	370864	370638	388782	380688
	Frequency (SSRs/Mb)	9.69	9.87	9.85	9.45
	Density (bp/Mb)	185.06	186.00	193.50	175.25
AAT	Number of SSRs	8810	8608	8720	8927
	Total length (bp)	241386	186990	259371	203265
	Frequency (SSRs/Mb)	4.40	4.32	4.34	4.11
	Density (bp/Mb)	120.45	93.84	129.09	93.58
AAC	Number of SSRs	7650	7541	7671	6680
	Total length (bp)	158211	145791	164925	121278
	Frequency (SSRs/Mb)	3.82	3.78	3.82	3.08
	Density (bp/Mb)	78.95	73.16	82.09	55.83
AAAT	Number of SSRs	17207	17157	17213	17377
	Total length (bp)	405036	345340	402548	354620
	Frequency (SSRs/Mb)	8.59	8.61	8.57	8.00
	Density (bp/Mb)	202.11	173.31	200.35	163.25
AAAC	Number of SSRs	17045	17937	17204	16339
	Total length (bp)	320028	331308	326264	297960
	Frequency (SSRs/Mb)	8.51	9.00	8.56	7.52
	Density (bp/Mb)	159.69	166.26	162.39	137.17
AAAG	Number of SSRs	10940	11640	11391	11413
	Total length (bp)	446300	346432	327312	340236
	Frequency (SSRs/Mb)	5.46	5.84	5.67	5.25
	Density (bp/Mb)	222.70	173.85	162.91	156.63
AAGG	Number of SSRs	7870	8538	8096	8167
	Total length (bp)	232180	219804	281628	196244
	Frequency (SSRs/Mb)	3.93	4.28	4.03	3.76
	Density (bp/Mb)	115.86	110.31	140.17	90.34
AATG	Number of SSRs	6953	6977	7016	7090
	Total length (bp)	137664	133672	136576	134172
	Frequency (SSRs/Mb)	3.47	3.50	3.49	3.26
	Density (bp/Mb)	68.69	67.08	67.98	61.77

Table 4 (continued)

Repeat motif type	Parameter	*C. dromedarius*	*C. bactrianus*	*C. ferus*	*V. pacos*
AGAT	Number of SSRs	5045	5072	5108	5371
	Total length (bp)	213992	158708	240380	163960
	Frequency (SSRs/Mb)	2.52	2.55	2.54	2.47
	Density (bp/Mb)	106.78	79.65	119.64	75.48
AAAAC	Number of SSRs	6646	6714	6766	6369
	Total length (bp)	163385	153615	162930	142350
	Frequency (SSRs/Mb)	3.32	3.37	3.37	2.93
	Density (bp/Mb)	81.53	77.09	81.09	65.53
AAAAT	Number of SSRs	2099	2114	2081	2145
	Total length (bp)	67650	56710	67885	57045
	Frequency (SSRs/Mb)	1.05	1.06	1.04	0.99
	Density (bp/Mb)	33.76	28.46	33.79	26.26
AAAAG	Number of SSRs	1887	2016	2057	1894
	Total length (bp)	66275	58070	86010	52595
	Frequency (SSRs/Mb)	0.94	1.01	1.02	0.87
	Density (bp/Mb)	33.07	29.14	42.81	24.21
AAAAAC	Number of SSRs	1554	1651	1626	1399
	Total length (bp)	46200	44994	48330	36954
	Frequency (SSRs/Mb)	0.78	0.83	0.81	0.64
	Density (bp/Mb)	23.05	22.58	24.05	17.01

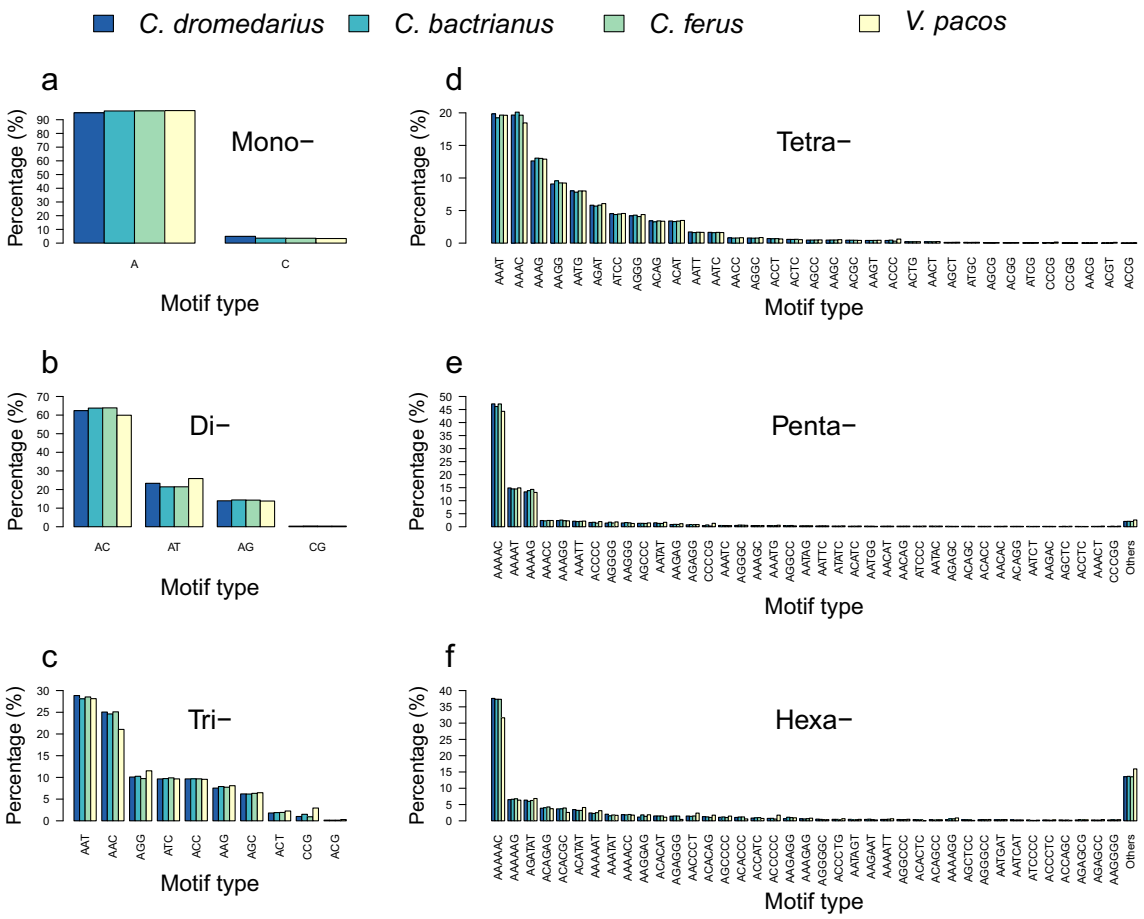

Fig. 3 Percentage of SSR motif types in the camelid genomes. Percentages were calculated according to the total number of each SSR motif type divided by the total number of SSRs for that SSR type in each genome. ABCDEF represent mono-, di-, tri-, tetra-, penta-, and hexanucleotide SSR types, respectively

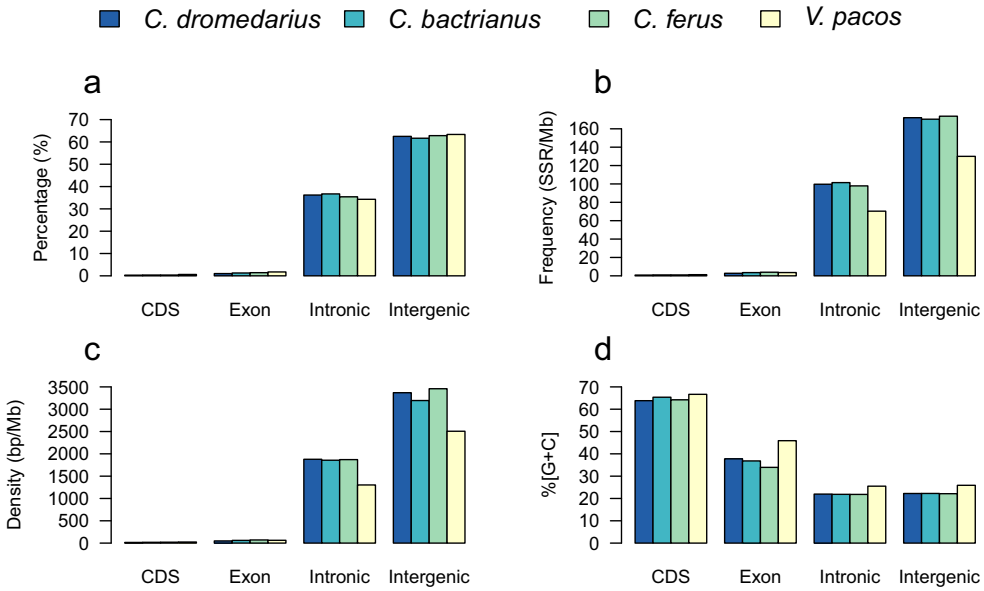

Fig. 4 Comparison of percentage, frequency, density, and GC content of SSRs in different genomic regions of the camelid species. ABCD represent percentage, frequency, density, and GC content of SSRs, respectively

1870.78, and 1302.66 bp/Mb, and intergenic regions of 3369.28, 3194.22, 3458.25, and 2505.98 bp/Mb (Fig. 4c).

In addition, the GC content of SSRs was investigated for different genomic regions of the four camelid genomes (Fig. 4d). GC contents were almost identical for *C. dromedarius*, *C. bactrianus*, *C. ferus*, and *V. pacos*. GC contents were found to vary between different genomic regions (Fig. 4d), but the distributions in intronic and intergenic regions were highly similar. SSRs located in CDSs were found to have the highest GC content (63.82–66.66%), followed by those in exons (33.94–45.89%), intronic regions (21.82–25.51%), and finally intergenic regions (22.14–25.90%).

In CDSs, trinucleotide SSRs were the most abundant type, followed by hexa-, mono-, tetra-, di-, and pentanucleotide SSRs (Fig. 5a). In exons, mononucleotide SSRs were the most abundant type in *C. dromedarius*, *C. bactrianus*, and *C. ferus*, while trinucleotide SSRs were the most abundant type in *V. pacos* (Fig. 5b). Hexanucleotide SSRs were the least abundant type in the exons of *C. bactrianus* and *C. ferus*, versus pentanucleotide SSRs in the exons of *C. dromedarius* and *V. pacos* (Fig. 5b). In intronic

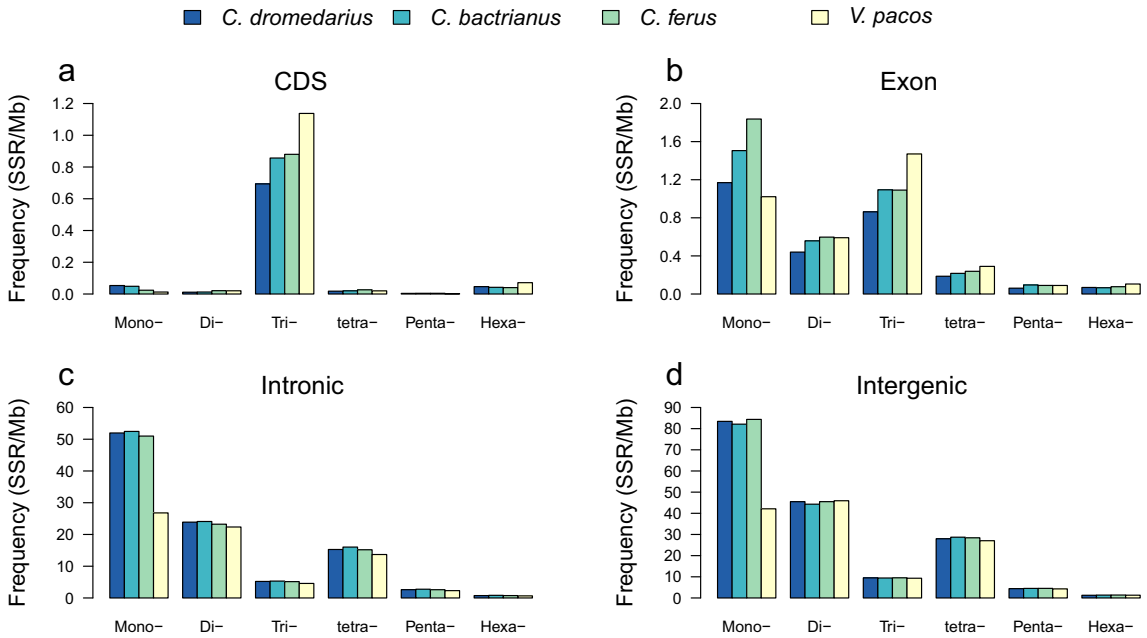

Fig. 5 Relative frequency of mono- to hexanucleotide SSRs in different genomic regions of the camelid genomes. ABCD represent CDSs, exons, intronic regions, and intergenic regions, respectively

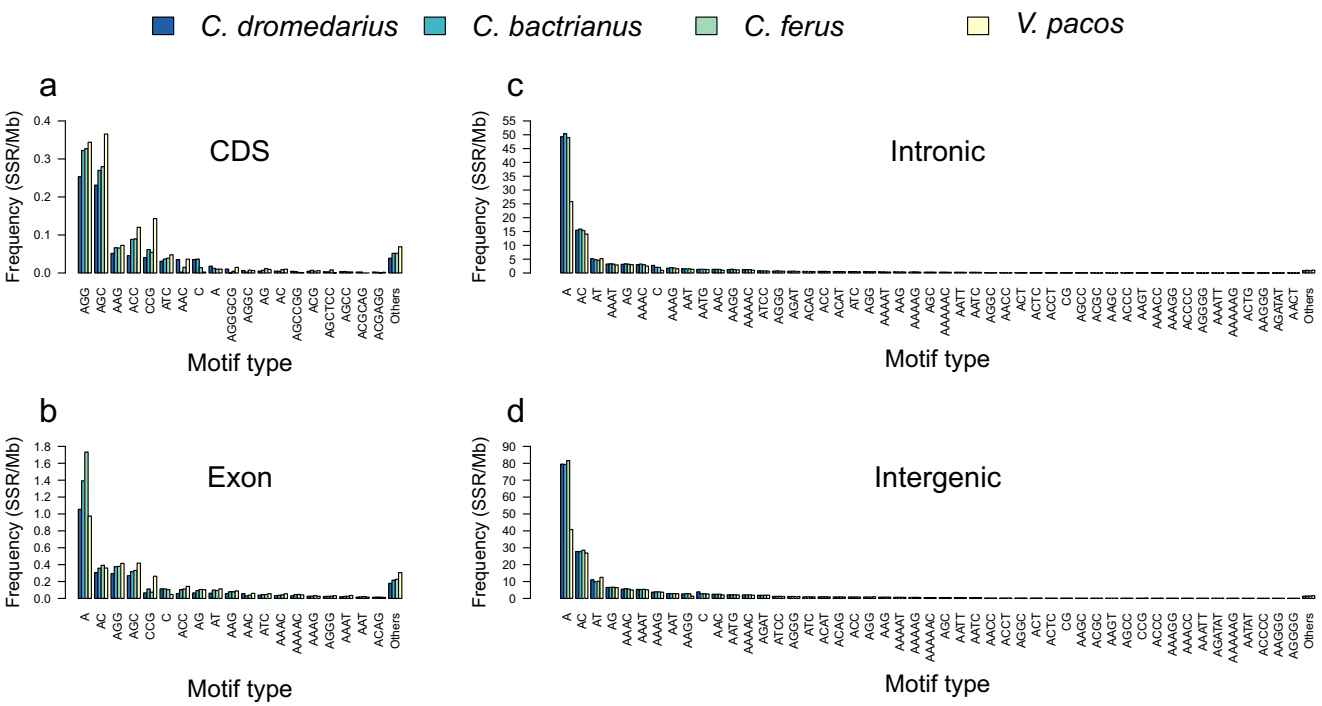

Fig. 6 Relative frequency of SSR motif types in different genomic regions of the camelid species. ABCD represent CDSs, exons, intronic regions, and intergenic regions, respectively

regions, mononucleotide SSRs were the most abundant type in all four camelid species, followed in decreasing order by di-, tetra-, tri-, penta-, and hexanucleotide SSRs (Fig. 4c). In intergenic regions, mononucleotide SSRs were the most abundant type in *Camelus* species, while dinucleotide SSRs were the most abundant type in *V. pacos* (Fig. 4d). Trinucleotide SSRs were rare in intergenic and intronic regions for all four camelid species, and hexanucleotide SSRs were the least abundant type in intronic and intergenic regions (Fig. 4c, d).

The abundances of specific repeat motif types were found to vary distinctly in different genomic regions of the four species (Fig. 6). In CDS regions, the predominant motif was $(AGG)_n$ in the three *Camelus* species, accounting for ~ 30% of CDS SSRs, followed by $(AGC)_n$ at ~ 28% (Fig. 6a). Meanwhile, $(AGC)_n$ was the most abundant trinucleotide repeat in the CDSs of *V. pacos*, followed by $(AGG)_n$; these together accounted for 56.14% of CDS SSRs. In all four genomes, the motifs $(AC)_n$, $(AGG)_n$, and $(AGC)_n$ had similar abundances in CDS regions, together accounting for 39.65–44.19% of CDS SSRs (Fig. 4b). Consistently, the $(A)_n$ motif was the most abundant repeat in exons (27.33–44.09%), intronic regions (36.65–50.02%), and intergenic regions (31.37–46.98%) (Fig. 4b, c, d). $(AC)_n$ was the second most frequent motif in intronic (15.54–19.95%) and intergenic regions(16.14–20.67%), followed by $(AT)_n$, which comprised 4.70–7.38% and 6.43–9.62% of the SSRs in intronic and intergenic regions, respectively (Fig. 4c, d).

Discussion

Diversity of microsatellite distribution in camelid genomes

In this study, microsatellites with motifs of 1–6 bp were identified using `PERF` with consistent search parameters in four camelid species (*C. dromedarius*, *C. bactrianus*, *C. ferus*, and *V. pacos*). The number of SSRs, relative frequency, relative density, and GC content were analyzed to understand the structure and diversity of SSR content in camelid genomes. The findings provide evidence that these four genomes have similar distribution patterns for SSRs, suggesting that other camelid genomes are likely to share the same pattern. However, our results showed that the SSR density did not drive the genome size in these four camelids. Instead, there was a negative correlation between SSR densities and genome sizes, suggesting that SSRs might have not contributed significantly to the expansion of the genome in evolution. Perfect SSRs were found to comprise 0.53% of the *C. dromedarius* and *C. ferus* genomes, 0.51% in *C. bactrianus*, and 0.38% in *V. pacos*. The total percentages of SSRs were higher in the three *Camelus* species than in bovids (0.44–0.48%) (Qi et al. 2015; Ma 2015), but lower than in macaques (0.83–0.88%) (Liu et al. 2017) and humans (3%) (Subramanian et al. 2002). The wide variance in total percentages may arise from the use of different computational methods for SSR

mining, the relative completeness of different genome assemblies, or real differences in SSR content among these species (Sharma et al. 2007).

As expected, the six types of SSRs were not evenly abundant across the four camelid genomes. Mononucleotide SSRs were the most abundant repeat type, consistent with bovids (Qi et al. 2015; Ma 2015) and macaques (Liu et al. 2017). In addition, this finding is consistent with the previous report that mononucleotide SSR repeats are more frequent in eukaryotic genomes than other SSR repeat types (Sharma et al. 2007). However, dinucleotide SSR repeats are the most frequent type in dicotyledons (Kumpatla and Mukhopadhyay 2005), *Taenia solium* (Pajuelo et al. 2015), Drosophila (Katti et al. 2001), and rodents (Toth 2000), while trinucleotide SSR repeats are the most prevalent type in a number of prokaryotes (Kim et al. 2008; Sharma et al. 2007) and yeast (Katti et al. 2001). The second most frequent SSRs in camelid genomes are dinucleotides, accounting for 25.08–33.94% of all SSRs. The third most abundant SSRs are tetranucleotides, followed by tri-, penta-, and hexanucleotide SSRs. In this analysis, hexanucleotide repeats were the least frequent, at less than 2.22 SSRs/Mb, and accounted for only 0.76–1.00% of the total number of SSRs. This observation in camelids is similar to what has been found in humans (Subramanian et al. 2002), bovids (Qi et al. 2015), and macaques (Liu et al. 2017).

A comparative analysis was conducted for microsatellite motifs within each type of repeat. We observed variation in overall number, frequency, and density between the four camelids. However, SSR motif occurrences are expected to increase as the motif length decreases, as seen in some other species (Karaoglu et al. 2004; Qi et al. 2015; Liu et al. 2017). The most prevalent SSR motifs for each type were found to be almost identical across the four genomes. Among mononucleotide repeats, the motif $(A/T)_n$ was the most abundant, accounting for 95.06–96.66% of mononucleotide SSRs. Conversely, the motif $(C/G)_n$ was rare. The $(A/T)_n$ motif is also predominant in *Volvariella volvacea*, *Agaricus bisporus*, *Coprinus cinereus* (Wang et al. 2014), and *Caenorhabditis elegans* (Castagnone-Sereno et al. 2010), while the $(C/G)_n$ motif is the most frequent in *Meloidogyne incognita*, *Pristionchus pacificus* (Castagnone-Sereno et al. 2010), and *Schizophyllum commune* (Wang et al. 2014). Among dinucleotide SSRs, the most abundant motif was $(AC)_n$, similar to the trend observed in Carlavirus (Alam et al. 2014), humans (Subramanian et al. 2002), bovids (Qi et al. 2015), and macaques (Liu et al. 2017). The second most frequent dinucleotide motif was $(AT)_n$, followed by $(AG)_n$ and $(CG)_n$ motifs, which is consistent with *Bos grunniens* (Ma 2015). The rareness of $(CG)_n$ motifs can be explained by the tendency to AT richness, and by the fact that strand separation is harder for CG than for AT and other tracts, raising the potential of slipped strand mispairing (Zhao et al. 2011). The $(AAT)_n$ motif was the most frequent trinucleotide SSR in the four camelids, similar to macaques (Liu et al. 2017), *P. pacificus*, *M. hapla*, *B. malayi* (Castagnone-Sereno et al. 2010), and *Ziziphus jujuba* (Xiao et al. 2015); $(AAT)_n$ is conversely rare in *P. ostreatus*, *Coprinus cinereus*, and *S. commune* (Wang et al. 2014). A previous study revealed that the $(AAAT)_n$ motif predominates in *Ailuropoda melanoleuca* (Huang et al. 2015). Among tetra-, penta-, and hexanucleotide motif types, AT-rich SSR motifs including $(AAAT)_n$, $(AAAAC)_n$, and $(AAAAAC)_n$ were found to be predominant, which is consistent with macaques (Liu et al. 2017). Interestingly, none of the most prevalent SSR motifs includes exclusively Cs or Gs. The over-represented motifs identified in this study support the conclusion that nucleotide sequences with higher GC content are expected to contain fewer SSRs than those of higher AT content (Schlötterer 1998). Overall, the great similarity of the most abundant motifs between the four camelids is a strong indication that the pattern of microsatellites is conserved in genus *Camelus*.

Diversity of microsatellite distribution in different genomic regions

Substantial evidence exists that the genomic distribution of SSRs is nonrandom, presumably due to their influences on processes such as chromatin organization, gene activity, DNA repair, and DNA recombination (Li et al. 2002, 2004). This may indicate that SSRs in different genomic regions play different functional roles. For instance, SSR expansions or contractions in coding regions can control gene activation, while SSRs located in intronic regions impact gene transcription or mRNA splicing (Li et al. 2004). SSRs in coding regions may affect phenotypes, causing neuronal diseases and cancers in humans (Pearson et al. 2005; Li et al. 2004). Furthermore, SSR repeat variations in 5′ UTRs may affect gene expression, and longer SSR repeats located in 3′ UTRs may lead to transcription slippage (Li et al. 2004). Here, we further studied the distribution of SSRs in different genomic regions of four camelids. The results revealed extensive variation in the distributional patterns of different SSR types between different genomic regions of camelids. Our results also demonstrated great similarity in SSR distributions within the same genomic regions of these camelid species. SSRs in noncoding regions were found to be more abundant than in coding regions, which confirm results previously reported in eukaryotes (Toth 2000; Katti et al. 2001; Qi et al. 2016) and plants (Morgante et al. 2002; Lawson and Zhang 2006; Hong et al. 2007). SSRs were most frequent in intergenic regions, followed in order by intronic regions, exons, and CDSs. SSR abundance was lowest in

CDS regions, consistent with selection against frameshift mutations in coding regions (Li et al. 2002).

In CDSs, trinucleotide SSRs were the most frequent type, consistent with results observed in primates (Qi et al. 2016) and bovids (Qi et al. 2018). Such predominance of triplets over other SSR repeat types in coding regions may be explained by purifying selection, which serves to eliminate non-trimeric SSRs in coding regions as they may cause frameshift mutations (Metzgar et al. 2000). This strong evolutionary pressure against SSR expansions in CDS regions may maintain the stability of the protein products (Dokholyan et al. 2000). Mononucleotide SSRs were the most abundant in exons, intronic, and intergenic regions, with the exception of *V. pacos*, in which trinucleotide and dinucleotide SSRs were identified to be most frequent types in exons and intergenic regions, respectively. This was consistent with observations from other eukaryotic genomes (Sharma et al. 2007; Qi et al. 2016; Qi et al. 2018). Pentanucleotide SSRs were the least common type in CDSs, whereas hexanucleotide SSRs were the least common type in exons and intronic and intergenic regions, except in *C. dromedarius* and *V. pacos*, where pentanucleotide SSRs were the least common type in exons. The paucity of trinucleotide SSRs compared to di- and tetranucleotide SSRs was also quite pronounced in intronic and intergenic regions of the four camelids. This might be a signature of selection removing triplet repeats from noncoding regions because they could generate false open reading frames (Gonthier et al. 2015).

Comparisons among different genomic regions in the four camelid genomes demonstrated that the major SSR motif types showed great similarity in their relative abundances. The nonrandom distribution of SSRs in different genomic regions shows bias to several specific repeat motifs, suggesting that SSRs of different types may play different roles in different genomic regions (Li et al. 2004; Gemayel et al. 2012). For instance, $(AGG)_n$ repeats are predominant in the coding regions of primates (Qi et al. 2016) and bovids (Qi et al. 2018). Consistent with those results, this study found $(AGG)_n$ repeats to be the most frequent motifs in CDS regions of camelid genomes, followed by $(AGC)_n$ repeats. $(AGG)_n$ and $(AGC)_n$ motifs were also more frequent in exonic regions, and relatively infrequent in intronic and intergenic regions. Trinucleotide and hexanucleotide repeats were more abundant in CDS regions than other motif types, consistent with previous reports (Borstnik 2002; Subramanian et al. 2003). Overall, $(A)_n$ repeats were the most abundant motifs in the exons, introns, and intergenic regions of these camelids, followed by dinucleotide $(AC)_n$ repeats; these trends are similar to findings in primates (Qi et al. 2016) and bovids (Qi et al. 2018). In addition, dinucleotide $(AT)_n$ and $(AG)_n$ repeats were relatively frequent in intronic and intergenic regions of the four camelid genomes. $(AAAT)_n$ and $(AAAC)_n$ motifs were comparatively more frequent than other tetranucleotide repeats in intronic and intergenic regions.

GC content and repeat number in different types of microsatellites

Previous studies reported a correlation between GC content and the genomic features of mammals, including methylation patterns, the distribution of repeat elements (Jabbari and Bernardi 1998), and gene density (Duret et al. 1994; Duret and Hurst 2001). A high level of GC content was found to be associated with gene expression (Ren et al. 2007) and DNA thermostability (Vinogradov 2003). GC-rich regions were also associated with many genes, suggesting a potential functional relevance for the distribution of GC content in mammals (Galtier et al. 2001). Microsatellite motifs with high GC content have been reported to cause some diseases in humans. For instance, a $(CGG)_n$ repeat exceeding 200 units in the 5′ untranslated region (UTR) of FMR1 was identified as the genetic cause of fragile X syndrome (Sharma et al. 2007). Furthermore, expansion of $(CGG)_n$ repeats in the 5′ UTR of the *DIP2B* gene causes FRA12A mental retardation (Winnepenninckx et al. 2007). $(G)_n$ repeats in the membrane protein gene pmp10 of *Chlamydophila* were reported to be involved in the virulence and pathogenesis of *Chlamydia* (Grimwood et al. 2001), and $(C)_n$ repeats in outer membrane proteins was found to be involved in the pathogenesis of *Clamydophila pneumoniae* (Rocha 2002). Additionally, high GC content may have significant roles in the entire viral genome. For example, G-string mutants in the thymidine kinase gene were found to be associated with reactivation of herpes simplex virus (Griffiths et al. 2006).

Our results revealed that GC content is remarkably consistent within a SSR type, and is not evenly distributed in different genomic regions. Our results also suggest that SSRs with high AT content are prevalent in each genome, similar to what has been reported in 26 eukaryotic genomes (Sharma et al. 2007). $(A/T)_n$ motifs were more predominant than $(G/C)_n$ motifs, which could be interpreted as being due to a high level of AT content in the majority of the analyzed SSRs. A previous study reported that trinucleotide SSRs have the highest GC content in bovids (Qi et al. 2015), which disagrees with our results. Here, dinucleotide SSRs were found to possess the highest GC content in camelid genomes, which is consistent with macaques (Liu et al. 2017). However, GC contents varied greatly among different genomic regions, with CDSs > exons > intronic regions > intergenic regions. The high level of GC content in coding regions was investigated to determine its relative influence on gene expression patterns. For example, the GC content of 5′ UTR has been found to be positively

correlated with gene expression in chickens (Rao et al. 2013). In addition, the high GC content in SSR motifs has been suggested to potentially impact genome structure. For instance, increasing $(CGG)_n$ repeats in the HSV-1 genome demonstrated considerable hairpin-forming and quadruplex-forming potential (Li et al. 2004).

A number of studies reported that SSR repeat count has an influence on gene expression. As an illustration, a promoter of *Saccharomyces cerevisiae* containing 25 tandem repeats of the $(CAG)_n$ motif allows expression of a *URA3* reporter gene and yields sensitivity to the drug 5-fluoroorotic acid, but expansion to 30 or more repeats turns off *URA3* and provides drug resistance (Miret et al. 1998). Promoter regions of *Escherichia coli* containing exactly 12 tandem repeats of the $(GAA)_n$ motif were found to express lac Z, while those with $(GAA)_{1216}$ and $(GAA)_{511}$ repeat motifs do not express lac Z (Liu et al. 2000). In this study, repeat lengths and maximum lengths were found to significantly differ within and between SSR repeat types among the four genomes. Notably, dramatically fewer SSRs were observed as the number of repeats increased. This observation can be explained by the effect of high mutation rates on longer repeats compared to shorter repeats within a given SSR type (Leopoldino and Pena 2002). In particular, SSR instability is suggested to increase as the stretch of the repeat motif increases. For instance, an *in vitro* study in human colorectal cells demonstrated that replication error in a $(G)_{16}$ repeat was 30-fold higher than for $(G)_{10}$, and in a $(CA)_{26}$ repeat were 10-fold higher than for $(CA)_{13}$ (Campregher et al. 2010). Overall, the GC content and repeat counts of SSRs may play significant roles in most species.

Conclusion

The current work has contributed to a detailed characterization of microsatellites in camelid genomes. The camelid genomes are predominated by AT-rich SSRs, and SSRs are nonrandomly distributed. Mononucleotide SSRs were the most frequent type, followed in order by di-, tetra-, tri-, penta-, and hexanucleotide SSRs. The greatest GC content was in dinucleotide SSRs and the least in mononucleotide SSRs. The number of SSRs, relative frequency, and relative density were generally found to decrease in these genomes as motif repeat length increased. SSRs were demonstrated to be more frequent in noncoding regions than in coding regions. Overall, the results of this study showed similar patterns of SSR distribution across the four camelid species, which indicates that the same pattern of microsatellites may apply to other camels. These data provide a comprehensive view into SSR genomic distribution in the Camelidae family. Such an understanding of the characteristics of microsatellites in camelid genomes will serve many useful purposes such as the development of camelids-specific genetic markers with broad applications, in particular for STR-based genotyping, paternity testing and molecular breeding.

Acknowledgments The authors would like to thank Casey M. Bergman at the Department of Genetics and Institute of Bioinformatics, University of Georgia, for reviewing the manuscript and giving valuable suggestions throughout this work.

Author contributions MMM and MBA conceived and designed the experiments; MMM, ATA, SNA, BMA, MAI, and SAB carried out the experiments; MMM, ATA, SNA, and BMA analyzed the data; MMM and MBA wrote the manuscript. All authors reviewed the manuscript.

Funding This work was funded by the Life Science and Environment Research Institute and the Center of Excellence for Genomics (grant 20-0078), King Abdulaziz City for Science and Technology, Saudi Arabia.

Compliance with ethical standards

Competing interests The authors declare that they have no competing interests.

References

Al-Swailem AM, Shehata MM, Abu-Duhier FM, Al-Yamani EJ, Al-Busadah KA, Al-Arawi MS, Al-Khider AY, Al-Muhaimeed AN, Al-Qahtani FH, Manee MM, Al-Shomrani BM, Al-Qhtani SM, Al-Harthi AS, Akdemir KC, Inan MS, Otu HH (2010) Sequencing, analysis, and annotation of expressed sequence tags for *Camelus dromedarius*. PLoS ONE 5:e10720

Alam CM, Singh AK, Sharfuddin C, Ali S (2014) Genome-wide scan for analysis of simple and imperfect microsatellites in diverse carlaviruses. Infect Genet Evol 21:287–294

Avvaru AK, Sowpati DT, Mishra RK (2017) PERF: an exhaustive algorithm for ultra-fast and efficient identification of microsatellites from large DNA sequences. Bioinformatics 27:573

Backes G, Hatz B, Jahoor A, Fischbeck G (2003) RFLP diversity within and between major groups of barley in Europe. Plant Breed 122:291–299

Bactrian Camels Genome Sequencing and Analysis Consortium, Jirimutu, Wang Z, Ding G, Chen G, Sun Y, Sun Z, Zhang H, Wang L, Hasi S, Zhang Y, Li J, Shi Y, Xu Z, He C, Yu S, Li S, Zhang W, Batmunkh M, Ts B, Narenbatu, Unierhu, Bat-Ireedui S, Gao H, Baysgalan B, Li Q, Jia Z, Turigenbayila, Subudenggerile, Narenmanduhu, Wang Z, Wang J, Pan L, Chen Y, Ganerdene Y, Dabxilt, Erdemt, Altansha, Altansukh, Liu T, Cao M, Aruuntsever, Bayart, Hosblig, He F, Zha-ti A, Zheng G, Qiu F, Sun Z, Zhao L, Zhao W, Liu B, Li C, Chen Y, Tang X, Guo C, Liu W, Ming L, Temuulen, Cui A, Li Y, Gao J, Li J, Wurentaodi, Niu S, Sun T, Zhai Z, Zhang M, Chen C, Baldan T, Bayaer T, Li Y, Meng H (2012) Genome sequences of wild and domestic bactrian camels. Nature Commun 3:1202

Borstnik B (2002) Tandem repeats in protein coding regions of primate genes. Genome Res 12:909–915

Campregher C, Scharl T, Nemeth M, Honeder C, Jascur T, Boland CR, Gasche C (2010) The nucleotide composition of microsatellites impacts both replication fidelity and mismatch repair in human colorectal cells. Zeitschrift für Gastroenterologie 48

Castagnone-Sereno P, Danchin EG, Deleury E, Guillemaud T, Malausa T, Abad P (2010) Genome-wide survey and analysis of microsatellites in nematodes, with a focus on the plant-parasitic species Meloidogyne incognita. BMC Genomics 11:598

Dokholyan NV, Buldyrev SV, Havlin S, Stanley HE (2000) Distributions of dimeric tandem repeats in non-coding and coding dna sequences. J Theoretical Biol 202:273–282

Duret L, Hurst LD (2001) The elevated GC content at exonic third sites is not evidence against neutralist models of isochore evolution. Mol Biol Evol 18:757–762

Duret L, Mouchiroud D, Gouy M (1994) HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res 22:2360–2365

Rocha EPC (2002) Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. Nucleic Acids Res 30:2031–2042

Galtier N, Piganeau G, Mouchiroud D, Duret L (2001) Gc-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907–911

Gemayel R, Cho J, Boeynaems S, Verstrepen KJ (2012) Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. Genes 3:461–480

Gonthier P, Sillo F, Lagostina E, Roccotelli A, Santa Cacciola O, Stenlid J, Garbelotto M (2015) Selection processes in simple sequence repeats suggest a correlation with their genomic location: insights from a fungal model system. BMC Genomics 16:1107

Griffiths A, Link MA, Furness CL, Coen DM (2006) Low-level expression and reversion both contribute to reactivation of herpes simplex virus drug-resistant mutants with mutations on homopolymeric sequences in thymidine kinase. J Virol 80:6568–6574

Grimwood J, Olinger L, Stephens RS (2001) Expression of Chlamydia pneumoniae polymorphic membrane protein family genes. Infect Immun 69:2383–2389

Groeneveld LF, Lenstra JA, Eding H, Toro MA, Scherf B, Pilling D, Negrini R, Finlay EK, Jianlin H, Groeneveld E, Weigend S, GLOBALDIV Consortium (2010) Genetic diversity in farm animals–a review. Animal Genetics 41 Suppl 1:6–31

Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y (2000) Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. Genome Res 10:62–71

Hancock J (1995) The contribution of slippage-like processes to genome evolution. J Mol Evol 41

Hong CP, Piao ZY, Kang TW, Batley J, Yang T, Hur Y, Bhak J, Park B, Edwards D, et al. (2007) Genomic distribution of simple sequence repeats in *Brassica rap*a. Molecules and Cells 23:349

Huang J, Li Y-Z, Du L-M, Yang B, Shen F-J, Zhang H-M, Zhang Z-H, Zhang X-Y, Yue B-S (2015) Genome-wide survey and analysis of microsatellites in giant panda *(Ailuropoda melanoleuca)*, with a focus on the applications of a novel microsatellite marker system. BMC Genomics 16:61

Huntley MA, Golding GB (2006) Selection and slippage creating serine homopolymers. Mol Biol Evol 23:2017–2025

Jabbari K, Bernardi G (1998) Cpg doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene 224:123–128

Jugran AK, Bhatt ID, Rawal RS, Nandi SK, Pande V (2013) Patterns of morphological and genetic diversity of *Valeriana jatamansi* Jones in different habitats and altitudinal range of West Himalaya, India. Flora - Morphology, Distribution. Functional Ecology of Plants 208:13–21

Jurka J, Pethiyagoda C (1995) Simple repetitive DNA sequences from primates: compilation and analysis. Journal of Molecular Evolution 40:120–126

Karaoglu H, Lee CMY, Meyer W (2004) Survey of simple sequence repeats in completed fungal genomes. Mol Biol Evol 22:639–649

Karaoglu H, Lee CMY, Meyer W (2005) Survey of simple sequence repeats in completed fungal genomes. Molecular Biology and Evolution 22:639–649

Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167

Kim T-S, Booth JG, Gauch HG, Sun Q, Park J, Lee Y-H, Lee K (2008) Simple sequence repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference. BMC Genomics 9:31

Kumpatla SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome 48:985–998

Last L, Lüscher G, Widmer F, Boller B, Kölliker R (2014) Indicators for genetic and phenotypic diversity of *Dactylis glomerata* in Swiss permanent grassland. Ecol Indic 38:181–191

Lawson MJ, Zhang L (2006) Distinct patterns of SSR distribution in the arabidopsis thaliana and rice genomes. Genome Biol 7:R14

Leopoldino AM, Pena SD (2002) The mutational spectrum of human autosomal tetranucleotide microsatellites. Hum Mutat 21:71–79

Li C-Y, Liu L, Yang J, Li J-B, Su Y, Zhang Y, Wang Y-Y, Zhu Y-Y (2009a) Genome-wide analysis of microsatellite sequence in seven filamentous fungi. Interdisciplinary Sciences: Computational Life Sciences 1:141–150

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009b) The sequence alignment/map format and SAMtools. Bioinformatics (Oxford, England) 25:2078–2079

Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Molecular Ecology 11:2453–2465

Li Y-C, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21:991–1007

Liu L, Dybvig K, Panangala VS, van Santen VL, French CT (2000) GAA trinucleotide repeat region regulates M9/pMGA gene expression in *Mycoplasma gallisepticum*. Infect Immun 68:871–876

Liu S, Hou W, Sun T, Xu Y, Li P, Yue B, Fan Z, Li J (2017) Genome-wide mining and comparative analysis of microsatellites in three macaque species. Mol Gen Genomics 292:537–550

Ma Z (2015) Genome-wide characterization of perfect microsatellites in yak *(Bos grunniens)*. Genetica 143:515–520

Manee MM, Alharbi SN, Algarni AT, Alghamdi WM, Altammami MA, Alkhrayef MN, Alnafjan BM (2017) Molecular cloning, bioinformatics analysis, and expression of small heat shock protein beta-1 from *Camelus dromedarius*, Arabian camel. PLOS ONE 12:e0189905

Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10:72–80

Miret JJ, Pessoa-Brandão L, Lahue RS (1998) Orientation-dependent and sequence-specific expansions of CTG/CAG trinucleotide repeats in *Saccharomyces cerevisiae*. Proceedings of the National Academy of Sciences 95:12438–12443

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Mrazek J, Guo X, Shah A (2007) Simple sequence repeats in prokaryotic genomes. Proceedings of the National Academy of Sciences 104:8472–8477

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. (2012) BEDOPS: high-performance genomic feature operations. Bioinformatics 28:1919–1920

Pajuelo MJ, Eguiluz M, Dahlstrom EW, Requena D, Guzmán F, Ramírez M, Sheen P, Frace M, Sammons SA, Cama VA, Anzick SL, Bruno D, Mahanty S, Wilkins PP, Nash TE, Gonzalez AE, García HH, Gilman RH, Porcella SF, Zimic M (2015) Identification and characterization of microsatellite markers derived from the whole genome analysis of *Taenia solium*. PLoS Neglected Tropical Diseases 9 12:e0004316

Pearson CE, Nichol Edamura K, Cleary JD (2005) Repeat instability: mechanisms of dynamic mutations. Nature reviews. Genetics 6:729–742

Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40:D130–D135

Qi W-H, Jiang X-M, Du L-M, Xiao G-S, Hu T-Z, Yue B-S, Quan Q-M (2015) Genome-wide survey and analysis of microsatellite sequences in bovid species. PLOS ONE 10:e0133667

Qi W-H, Jiang X-M, Yan C-C, Zhang W-Q, Xiao G-S, Yue B-S, Zhou C-Q (2018) Distribution patterns and variation analysis of simple sequence repeats in different genomic regions of bovid genomes. Scientific Reports 8:14407

Qi W-H, Yan C-C, Li W-J, Jiang X-M, Li G-Z, Zhang X-Y, Hu T-Z, Li J, Yue B-S (2016) Distinct patterns of simple sequence repeats and GC distribution in intragenic and intergenic regions of primate genomes. Aging 8:2635–2654

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842

Rao YS, Chai XW, Wang ZF, Nie QH, Zhang X (2013) Impact of GC content on gene expression pattern in chicken. Genetics Selection Evolution 45:9

Ren L, Gao G, Zhao D, Ding M, Luo J, Deng H (2007) Developmental stage related patterns of codon usage and genomic GC content: searching for evolutionary fingerprints with models of stem cell differentiation. Genome Biol 8:R35

Schlötterer C (1998) Genome evolution: are microsatellites really simple sequences? Curr Biol 8:R132–R134

Sharma PC, Grover A, Kahl G (2007) Mining microsatellites in eukaryotic genomes. Trends Biotechnol 25:490–498

Shehzad T, Okuizumi H, Kawase M, Okuno K (2009) Development of SSR-based sorghum (*Sorghum bicolor* (L.) Moench) diversity research set of germplasm and its evaluation by morphological traits. Genetic Resources and Crop Evolution 56:809–827

Subramanian S, Madgula VM, George R, Mishra RK, Pandit MW, Kumar CS, Singh L (2003) Triplet repeats in human genome: distribution and their association with genes and other genomic regions. Bioinformatics 19:549–552

Subramanian S, Mishra RK, Singh L (2002) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biology 4:R13–R13

Toth G (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10:967–981

Vinogradov AE (2003) DNA helix: the importance of being GC-rich. Nucleic Acids Res 31:1838–1844

Wang Y, Chen M, Wang H, Wang J-F, Bao D (2014) Microsatellites in the genome of the edible mushroom, *Volvariella volvacea*. BioMed Res Int 2014:1–10

Winnepenninckx B, Debacker K, Ramsay J, Smeets D, Smits A, FitzPatrick DR, Kooy RF (2007) CGG-repeat expansion in the DIP2b gene is associated with the fragile site FRA12a on chromosome 12q13.1. The American Journal of Human Genetics 80:221–231

Wu H, Guang X, Al-Fageeh MB, Cao J, Pan S, Zhou H, Zhang L, Abutarboush MH, Xing Y, Xie Z, Alshanqeeti AS, Zhang Y, Yao Q, Al-Shomrani BM, Zhang D, Li J, Manee MM, Yang Z, Yang L, Liu Y, Zhang J, Altammami MA, Wang S, Yu L, Zhang W, Liu S, Ba L, Liu C, Yang X, Meng F, Wang S, Li L, Li E, Li X, Wu K, Zhang S, Wang J, Yin Y, Yang H, Al-Swailem AM, Wang J (2014) Camelid genomes reveal evolution and adaptation to desert environments. Nat Commun 5:5188

Xiao J, Zhao J, Liu M, Liu P, Dai L, Zhao Z (2015) Genome-wide characterization of simple sequence repeat (SSR) loci in chinese jujube and jujube SSR primer transferability. PLOS ONE 10:e0127812

Xu Y, Hu Z, Wang C, Zhang X, Li J, Yue B (2016) Characterization of perfect microsatellite based on genome-wide and chromosome level in Rhesus monkey *(Macaca mulatta)*. Gene 592:269–275

Yang J, Wang J, Chen L, Yu J, Dong J, Yao Z-J, Shen Y, Jin Q, Chen R (2003) Identification and characterization of simple sequence repeats in the genomes of Shigella species. Gene 322:85–92

Zhao X, Tan Z, Feng H, Yang R, Li M, Jiang J, Shen G, Yu R (2011) Microsatellites in different Potyvirus genomes: survey and analysis. Gene 488:52–56