SCIENTIFIC REPORTS

OPEN

Distribution patterns and variation analysis of simple sequence repeats in different genomic regions of bovid genomes

Received: 26 February 2018
Accepted: 4 September 2018
Published online: 26 September 2018

Wen-Hua Qi[1], Xue-Mei Jiang[2], Chao-Chao Yan[3], Wan-Qing Zhang[4], Guo-Sheng Xiao[1], Bi-Song Yue[3] & Cai-Quan Zhou[5]

As the first examination of distribution, guanine-cytosine (GC) pattern, and variation analysis of microsatellites (SSRs) in different genomic regions of six bovid species, SSRs displayed nonrandomly distribution in different regions. SSR abundances are much higher in the introns, transposable elements (TEs), and intergenic regions compared to the 3′-untranslated regions (3′UTRs), 5′UTRs and coding regions. Trinucleotide perfect SSRs (P-SSRs) were the most frequent in the coding regions, whereas, mononucleotide P-SSRs were the most in the introns, 3′UTRs, TEs, and intergenic regions. Trifold P-SSRs had more GC-contents in the 5′UTRs and coding regions than that in the introns, 3′UTRs, TEs, and intergenic regions, whereas mononucleotide P-SSRs had the least GC-contents in all genomic regions. The repeat copy numbers (RCN) of the same mono- to hexanucleotide P-SSRs showed significantly different distributions in different regions ($P < 0.01$). Except for the coding regions, mononucleotide P-SSRs had the most RCNs, followed by the pattern: di- > tri- > tetra- > penta- > hexanucleotide P-SSRs in the same regions. The analysis of coefficient of variability (CV) of SSRs showed that the CV variations of RCN of the same mono- to hexanucleotide SSRs were relative higher in the intronic and intergenic regions, followed by the CV variation of RCN in the TEs, and the relative lower was in the 5′UTRs, 3′UTRs, and coding regions. Wide SSR analysis of different genomic regions has helped to reveal biological significances of their distributions.

Microsatellites (or Simple sequence repeats, SSRs) are composed of tandem repeats of 1–6 oligonucleotides. It has been reported that SSRs play an important role in chromatin fractions, gene expression and regulation, as well as transcription and protein function[1,2]. They are hypermutable loci due to strand slippage and unequal recombination lead to indels of repeat units[3], which affect local structure of the DNA or protein sequences[2]. Variation of intronic SSRs can affect gene transcription and mRNA splicing[4]; Trinucleotide SSRs located in the UTRs (untranslated regions) or introns can also induce gene silencing[4]. Distribution of SSRs in the coding regions, 5′UTRs, introns, and 3′UTRs of genes are widely belived to affect transcription and translation as well as gene function[4]. The increase and decrease of SSR motifs in the 5′UTRs are known to regulate multiple characteristics[5–7]. Tri- and hexanucleotide SSRs in genes encode into amino acid, which may play particular roles in protein structure[8,9]. SSRs in both coding and regulatory regions can alter the structure of proteins or DNA when they expand beyond a certain length[10].

In silico mining and analysis of SSRs could help to disclose different aspects of the distribution and dynamics of SSRs in eukaryotic genomes[11]. There are two SSR search methods: using a suitable search tool (MISA[12], SciRoKo[13], msatcommander[14], GMATA[15], Krait[16]) and accessing a relevant SSR database (MMDBJ, SSRD, TRBase, InSatdb, and TRDB)[11]. The mining and analysis of SSRs not only helps in addressing biological questions,

[1]College of Biology and Food Engineering, Chongqing Three Gorges University, Chongqing, 404100, P. R. China. [2]College of Environmental and Chemistry Engineering, Chongqing Three Gorges University, Chongqing, 404100, P. R. China. [3]Key Laboratory of Bio-resources and Eco-environment (Ministry of Education), College of Life Sciences, Sichuan University, Chengdu, 610064, P. R. China. [4]College of Life Sciences, Sichuan Agricultural University, Ya'an, Sichuan Province, 625014, P. R. China. [5]Key Laboratory of Southwest China Wildlife Resources Conservation (Ministry of Education), China West Normal University, Nanchong, 637009, P. R. China. Correspondence and requests for materials should be addressed to C.-Q.Z. (email: drcqzhou1@163.com or 604598775@qq.com)

but also facilitates better utilizing of SSRs for multiple utilizations. The genome sequence data from six bovid species: *Bos taurus*, *Bos mutus*, *Bubalus bubalis*, *Ovis aries*, *Capra hircus*, and *Pantholops hodgsonii*, were used in this study. We detected and characterized SSRs and their motifs, and surveyed their distributions and variations in intragenic (i.e., 5′UTRs, coding regions, introns, and 3′UTRs) and intergenic regions. Furthermore, we addressed the questions of whether the abundance of different SSR types and motifs are similar or not in different genomic regions and how GC-content of SSR differ in 5′UTRs, coding regions, introns, 3′UTRs, transposable elements (TEs, or transposon), and intergenic regions. This research may facilitate our insight into SSR distribution of different genomic regions in the whole genome and GC-content difference of mono- to hexanucleotide SSRs. Repeat copy number (RCN) can provide some markers for studying processes of mutation and selection. Intragenic- and intergenic-wide analysis of SSR sequences of different bovid species has also improved our understanding of biological significances of SSR distributions.

Results

Distribution of SSRs in different genomic regions of bovid genomes. In the 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions of these bovids, P-SSRs was the most frequent type, and the least was in the complex SSRs (CX-SSRs, Fig. S1); the intronic and intergenic regions had the most abundant P-SSRs, followed by the pattern: 3′UTRs > 5′UTRs > TEs > coding regions (Fig. S1). The relative abundance of the same SSR types showed greatly similar in the same regions of bovid species.

In the 5′UTRs, tri- and mononucleotide P-SSRs were the most frequent type, followed by the pattern: di- > tetra- > penta- > hexanucleotide P-SSRs in the six bovid species (Fig. 1A and Table S1). In the coding regions, trinucleotide P-SSRs was the most frequent type, followed by the pattern: mono- > hexa- > di- > tetra- > pentanucleotide P-SSRs in these bovid species (Fig. 1B and Table S2). Pentanucleotide P-SSRs were relatively less frequent in the coding regions of these bovid species. In the 3′UTRs, mononucleotide P-SSRs was the most frequent type, followed by the pattern: di- > tri- > tetra- > penta- > hexanucleotide P-SSRs, the least was in the hexanucleotide P-SSRs in these species (Fig. 1D and Table S4). In the TEs, mononucleotide P-SSRs was the most frequent type, followed by the pattern: di- > tetra- > tri- > penta- > hexanucleotide P-SSRs in the bovid genomes (Fig. 1E and Table S5). In the TEs, mononucleotide P-SSRs was more than three times as frequent as di- and tetranucleotide P-SSRs, and interestingly, the latter are much more frequent than trinucleotide P-SSRs. In the intronic and intergenic regions, mononucleotide P-SSRs was the most frequent type, followed by the pattern: di- > tri- > penta- > tetra- > hexanucleotide P-SSRs, the least was in the hexanucleotide P-SSRs in these bovid species (Fig. 1C, F, and Tables S3, S6). In the introns, mononucleotide P-SSRs were more than twofold as frequent as dinucleotide P-SSRs. Interestingly, in the intronic and intergenic regions pentanucleotide P-SSRs are much more frequent than tetranucleotide P-SSRs, and hexanucleotide P-SSRs were relatively less abundant.

A comparison among these regions shows that relative abundance of the same mono- to hexanucleotide P-SSRs showed great similarity in the same genomic regions of these bovid species. Remarkably, the total SSR abundance among all regions for these species is the most for the intergenic regions (Fig. 2). There are more than five times the difference between the total SSR abundance of the coding regions and intergenic regions. SSR distribution seems to be the similarity between intronic and intergenic regions of these bovid genomes. These results here indicated that SSRs are more frequent in non-coding regions than coding regions in these bovid species.

Diversity of P-SSRs motifs in different genomic regions of bovid genomes. The abundance of different repeat motifs varied obviously with genomic regions in the six bovid species. In the 5′UTRs, the $(A)_n$ was the most frequent motif, followed by the motif $(CCG)_n$, thirdly the $(AGG)_n$, $(AC)_n$, and $(AG)_n$, fourthly the $(AGC)_n$ and $(ACG)_n$ (Fig. 2A). In the coding regions, the $(AGG)_n$ was the most frequent unit, followed by the motif $(ACG)_n$, $(AGC)_n$, and $(CCG)_n$, thirdly the $(ACC)_n$, $(AAG)_n$, $(A)_n$, and $(ACT)_n$ (Fig. 2B). In the introns, the $(A)_n$ was the most frequent unit, followed by the motif $(AC)_n$, thirdly the $(ACG)_n$, $(AGC)_n$, and $(AT)_n$, fourthly the $(AG)_n$, $(C)_n$, $(AAC)_n$, $(AAAT)_n$, and $(AAAC)_n$, the $(CG)_n$ and $(CCG)_n$ were relatively infrequent in the intronic regions (Fig. 2C). In the 3′UTRs, the $(A)_n$ was the most frequent motif, followed by the motif $(AC)_n$, thirdly the $(AT)_n$, fourthly the $(AG)_n$ and $(C)_n$ (Fig. 2D). In the TEs, the $(A)_n$ was the most frequent motif, followed by the motif $(AC)_n$ and $(AT)_n$, thirdly the $(AG)_n$ and $(AAAT)_n$, fourthly the $(C)_n$, $(AAT)_n$, $(AAC)_n$, $(AGC)_n$, and $(AAAC)_n$ (Fig. 2E). In the intergenic regions, the $(A)_n$ was the most frequent motif, followed by the motif $(AC)_n$, thirdly the $(AT)_n$, $(AGC)_n$, and $(ACG)_n$, fourthly the $(AG)_n$, $(C)_n$, $(AAAT)_n$, $(AAC)_n$, and $(AAAC)_n$ (Fig. 2F). Therefore, the motifs of SSRs are not randomly distributed in the 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions. There is a noticeable excess of $(CCG)_n$ repeat units in the 5′UTRs and coding regions compared to the introns, 3′UTRs, TEs, and intergenic regions. The $(AGG)_n$ repeat unit is obvious relatively abundant in the 5′UTRs and coding regions compared to other four regions. The $(ACG)_n$ and $(AGC)_n$ repeat units are relatively less abundant in the TEs compared to other five regions. The $(A)_n$ motif was significantly more frequent than the $(C)_n$ unit in the 5′UTRs, introns, 3′UTRs, TEs, and intergenic regions. The $(AAT)_n$ and $(AAC)_n$ units are relatively frequent in the TEs, where their abundance exceeds that of other trinucleotide motifs, and the $(CG)_n$ and $(CCG)_n$ motifs are relatively infrequent in the introns, TEs, 3′UTRs, and intergenic regions.

The GC-content of P-SSRs in different genomic regions of bovid genomes. The GC-content varied greatly among different genomic regions, but, in the same regions, the distribution of the GC-content is greatly similar. From the results (Fig. 3), we can know that 5′UTRs had the most GC-content (ranging 53.75–61.31%), followed by the coding regions (51.09–53.60%), next the 3′UTRs (42.61–45.18%) and TEs (42.53–42.83%), the least was the intronic (40.87–42.91%) and intergenic regions (41.39–41.84%). The distribution patterns of AT-contents (adenine-thymine content) showed greatly similar in the same genomic regions of these bovids (Table S7). From this we can know, high GC-content was distributed in exon-rich regions more frequently than other regions.

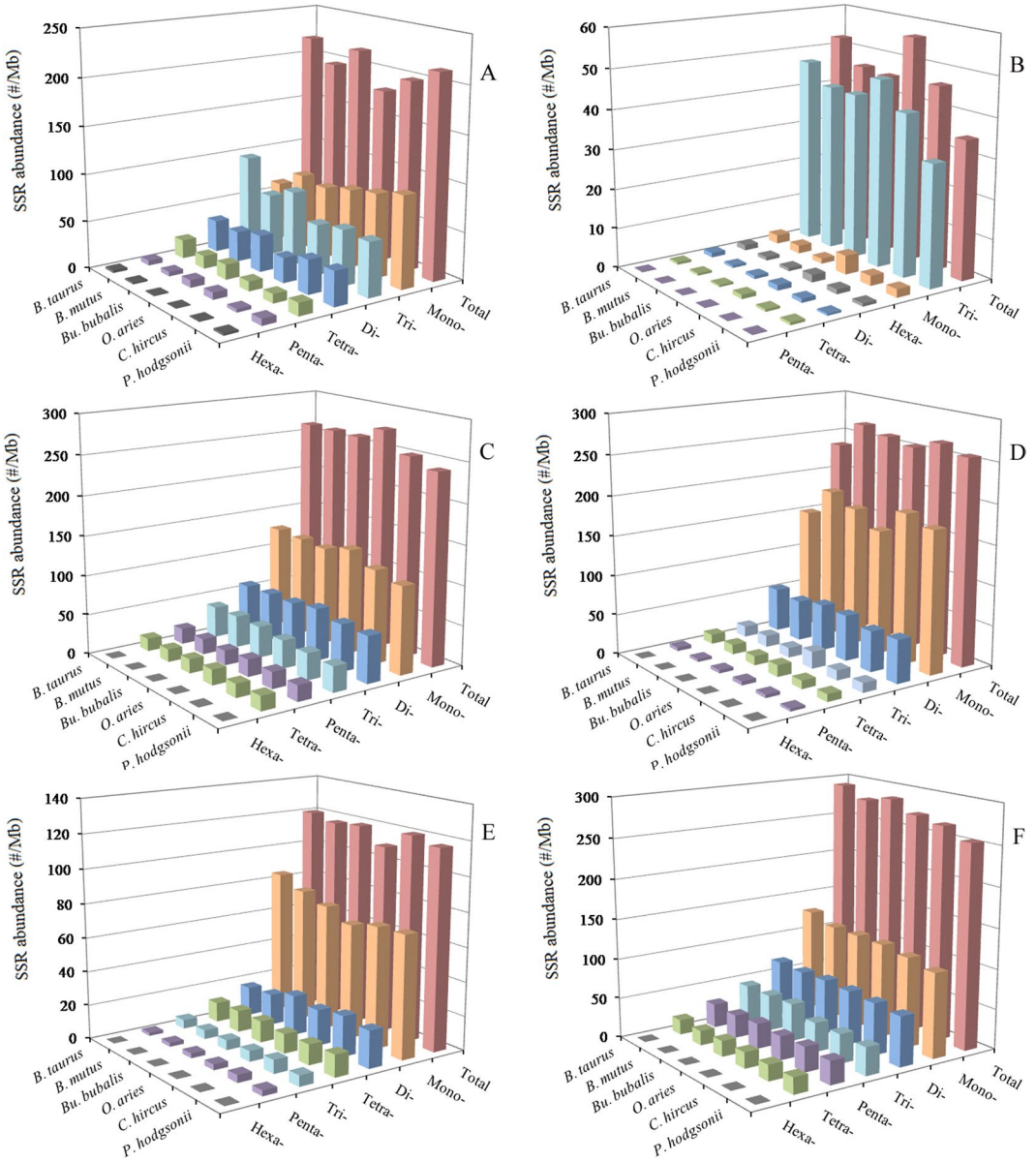

Figure 1. Relative abundance of mono- to hexanucleotide P-SSRs in different intragenic and intergenic regions of six bovids. ABCDEF represent 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions, respectively.

The AT- and GC-content of P-SSRs were calculated in the 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions of six bovid species, which the results were shown in Fig. 4 and Tables S8–13. In the six genomic regions, mononucleotide P-SSRs had the least GC-contents and were significantly less than their total GC-contents in these bovid genomes. In the 5′UTRs, except for the mononucleotide P-SSRs, the GC-content of the remaining nucleotide motifs are more than their AT-content (Fig. 4A and Table S8). Trinucleotide P-SSRs had the most GC-content (79.49–86.15%), followed by the pattern: hexa- > penta- (and tetra-) > di- > mononucleotide P-SSRs in the 5′UTRs of these bovid species (Fig. 4A). In contrast, the GC-content in the tri-, tetra- and hexanucleotide P-SSRs were more than their total GC-content in the 5′UTRs of these bovids (Fig. 4A). In the coding regions, the most GC-contents were in penta- and hexanucleotide P-SSRs, ranging from 68.00% (*P. hodgsonii*) to 92.80% (*B. taurus*), which were more than their AT-contents, and the GC-contents of mono-, di-, and tetranucleotide repeat types were significantly lower than their total GC-contents (61.67–70.58%) in these bovids, especially in mononucleotide P-SSRs (Fig. 4B and Table S9). In the 3′UTRs, except for the hexanucleotide P-SSRs, the GC-contents of the remaining nucleotide repeat units were less than their AT-contents, and mononucleotide P-SSRs had the least GC-contents (Fig. 4D and Table S11). In the intronic and intergenic regions, the most GC-contents were all in trinucleotide P-SSRs, followed by the pattern: penta- (and hexa-) > di- > tetra- > mononucleotide P-SSRs, and di-, penta-, and hexanucleotide P-SSRs are of similar GC-contents in the bovids (Fig. 4C, F and Tables S10, S13). In the TEs, we can know that the GC-contents of mono- to hexanucleotide P-SSRs are

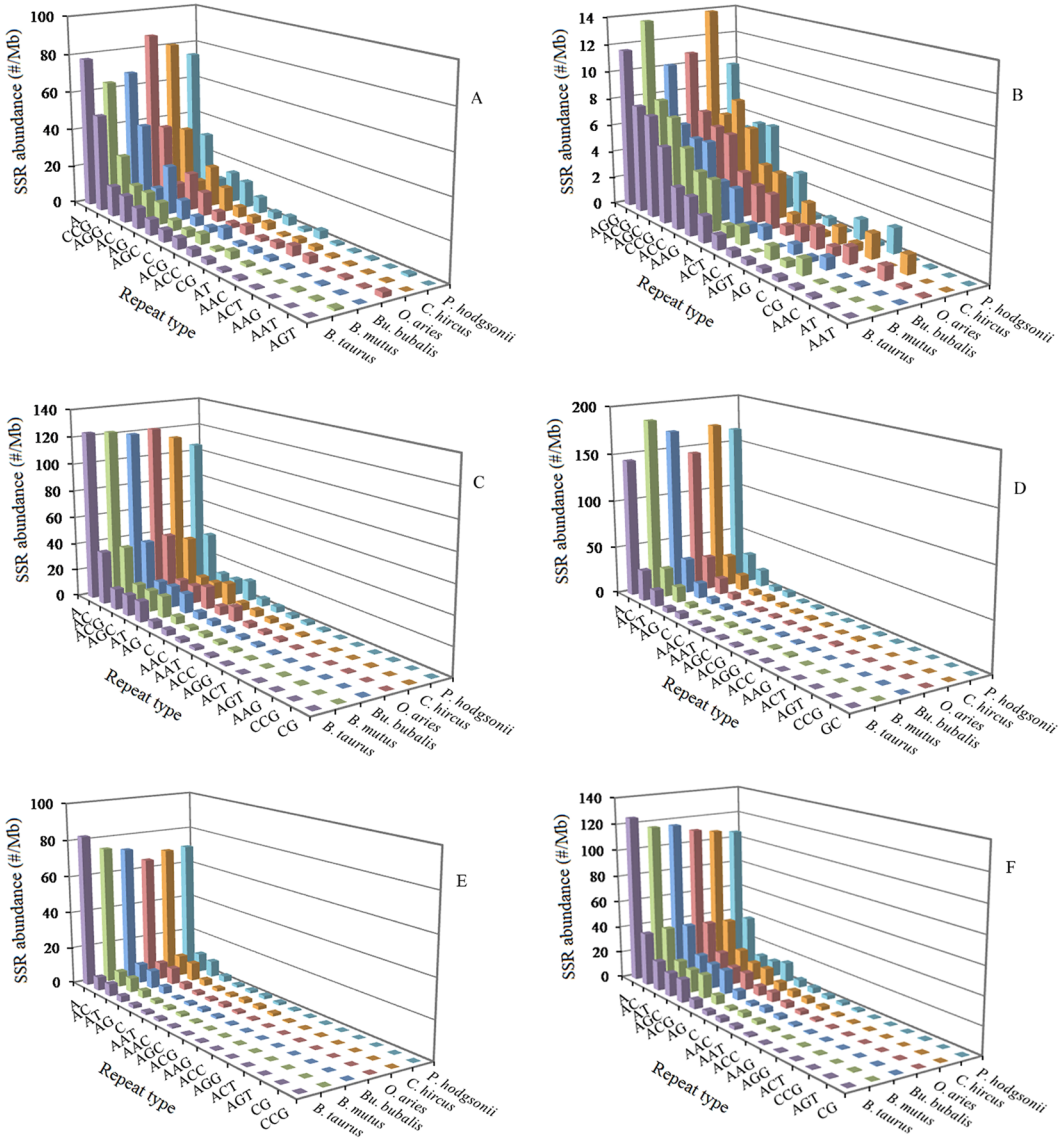

Figure 2. Distribution of different motfis of mono- to trinucleotide P-SSRs in different genomic regions of six bovid genomes. ABCDEF represent 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions, respectively.

less than their AT-contents, and the most GC-contents were all in tri- and hexanucleotide P-SSRs, followed by the pattern: di- (and penta-) > tetra- > mononucleotide P-SSRs, di- and pentanucleotide P-SSRs are of similar GC-contents in these bovids (Fig. 4E and Table S12). In contrast, the GC-contents of di- to hexanucleotide P-SSRs were more than their total GC-contents in the 3′UTRs and TEs, and the GC-contents of di-, tri-, penta-, and hexanucleotide P-SSRs were also more than their total GC-contents in the intronic and intergenic regions. In the 3′UTRs, introns, TEs, and intergenoic regions, their total AT-contents ranged from 71.20% to 89.29%, were obviously higher than their total GC-contents; whereas, in the coding regions, their total GC-contents ranged from 61.67% to 70.58%, were obviously higher than their total AT-contents in the bovids. Therefore, the GC-content of P-SSRs is probably high in coding-rich regions, whereas, the AT-content of P-SSRs is probably quite high in non-coding regions of these bovids.

The analysis of coefficient of variability (CV) of SSRs. The repeat copy numbers (RCN) of the same mono- to hexanucleotide SSRs had significantly different distributions in the different regions of these bovid genomes. The RCN of mono- and dinucleotide SSRs exhitbited great similar distributions and had the most counts of SSR loci in the intronic and intergenic regions, which were mainly distributed from 12 to 65 times and from 7 to 60 times, respectively (Fig. 5C, F and Fig. 6C, F). The RCN of mono- and dinucleotide SSRs were distributed from 10 to 60 times in the intronic and intergenic regions, which were clustered together and overlapped each other (Fig. 5C, F and Fig. 6C, F). The RCN of mono- and dinucleotide SSRs had the second most counts of SSR loci in the TEs, which were mainly distributed from 12 to 50 times and from 7 to 30 times, respectively. The

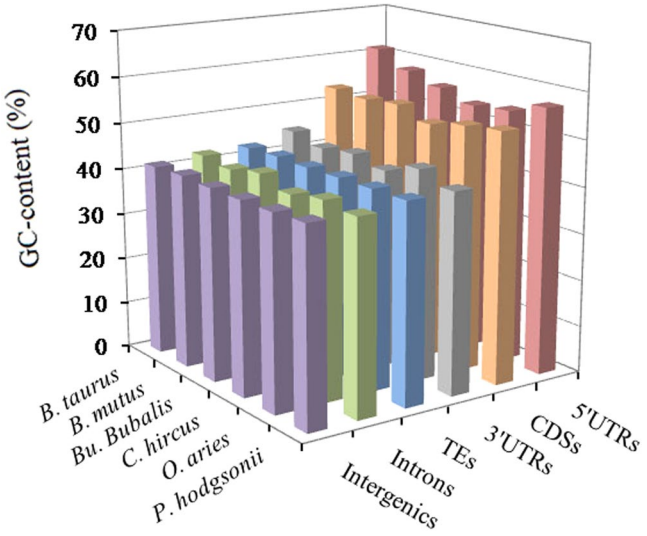

Figure 3. GC-contents of different intragenic and intergenic regions in six bovid species.

RCN of mononucleotide SSRs were distributed from 12 to 40 times in the TEs, which were clustered together and overlapped each other (Fig. 5C, F and Fig. 6C, F). In the 3′UTR regions, the RCN of mono- and dinucleotide SSRs displayed great similar among different bovid species, which were mainly distributed from 12 to 40 times and from 7 to 30 times, respectively (Fig. 5D and Fig. 6D). The RCN of mono- and dinucleotide SSRs had the fewest counts of SSR loci in the 5′UTRs and coding regions, which were mainly distributed from 12 to 30 times and from 7 to 20 times, respectively (Fig. 5A, B and Fig. 6A, B). The RCN of trinucleotide SSRs also showed great similar distributions and had the most counts of SSR loci in the intronic and intergenic regions, which were all mainly distributed from 5 to 40 times. The RCN of trinucleotide SSRs were distributed from 5 to 20 times in the intronic and intergenic regions, which were clustered together and overlapped each other (Fig. 7C, F). The RCN of trinucleotide SSRs had second most counts of SSR loci in the TEs, which were mainly distributed from 5 to 20 times (Fig. 7E). The RCN of trinucleotide SSRs had the fewest counts of SSR loci in the 5′UTRs, coding regions, and 3′UTRs, which were mainly distributed from 5 to 12 times (Fig. 7A, B, D).

The RCN of tetra- and pentanucleotide SSRs had most counts of SSR loci in the intronic and intergenic regions, which were mainly distributed from 4 to 30 times (Fig. S2C, F and Fig. S3C, F). The RCN of tetra- and pentanucleotide SSRs also showed great similar distributions and had second most counts of SSR loci in the TEs, which were mainly distributed from 4 to 12 times (Fig. S2E and Fig. S3E). The RCN of tetra- and pentanucleotide SSRs had fewer counts of SSR loci in the 5′UTR and 3′UTR regions, which were all mainly distributed from 4 to 6 times (Fig. S2A, D and Fig. S3A, D). The RCN of tetra- and pentanucleotide SSRs had fewest counts of SSR loci in the coding regions, which were mainly distributed from 4 to 5 times (Fig. S2B and Fig. S3B). The RCN of hexanucleotide SSRs had most counts of SSR loci in the intronic and intergenic regions, which were mainly distributed from 4 to 15 times (Fig. S4C, F). The RCN of hexanucleotide SSRs had second most counts of SSR loci in the TEs, which were mainly distributed from 4 to 9 times (Fig. S4E). The RCN of hexanucleotide SSRs were usually less and had fewer counts of SSR loci in the 5′UTRs, 3′UTRs, and coding regions, which were mainly distributed from 4 to 6 times (Fig. S4B).

The analysis of coefficient of variability (CV) of SSRs showed that the RCN of mono- and dinucleotide SSRs had relative higher variation in the 5′UTRs, 3′UTRs, TEs, introns, and intergenic regions of the same bovid species, followed by the CV pattern of RCN: trinucleotide SSRs > tetranucleotide SSRs > pentanucleotide SSRs > hexanucleotide SSRs (Fig. 8). In the coding regions, the RCN of mono- to trinucleotide SSRs had relative higher variation, followed by the CV pattern of RCN: hexanucleotide SSRs > tetranucleotide SSRs > pentanucleotide SSRs (Fig. 8). The CV variations of the same mono- to hexanucleotide SSRs showed a great deal of similarity in the 5′UTRs, 3′UTRs, and coding regions of these bovid genomes, which also showed similar in the intronic and intergenic regions, whereas they are slightly different from the CV variations of the same SSRs in the TEs (Fig. 8). The CV variations of RCN of the same mono- to hexanucleotide SSRs were relative higher in the intronic and intergenic regions, followed by the CV variation of RCN in the TEs, and the relative lower was in the 5′UTRs, 3′UTRs, and coding regions (Fig. 8). It has been inferred that SSR mutational rates within genes are inconsistent with those for SSRs located in other genomic regions.

Discussion

Similarity and diversity of P-SSR motifs in different genomic regions. It was presumed that SSR motifs were not distributed randomly in the different genomic regions and motif types may play important roles in gene expression and regulation[17–20]. The presence of SSRs in different genomic regions shows bias to some specific nucleotide motifs. The motifs of mono- to hexanucleotide P-SSR types showed distinct distributional patterns in the intragenic and intergenic regions of bovid species. In *drosophila*, coding regions exhibit a very high bias to $(AGC)_n$, and very rare for $(TGC)_n$[21]. In the study, there is also a noticeable excess of $(AGG)_n$ repeat units, and the second most frequent units are constituted by the $(ACG)_n$, $(AGC)_n$, and $(CCG)_n$ in the coding regions

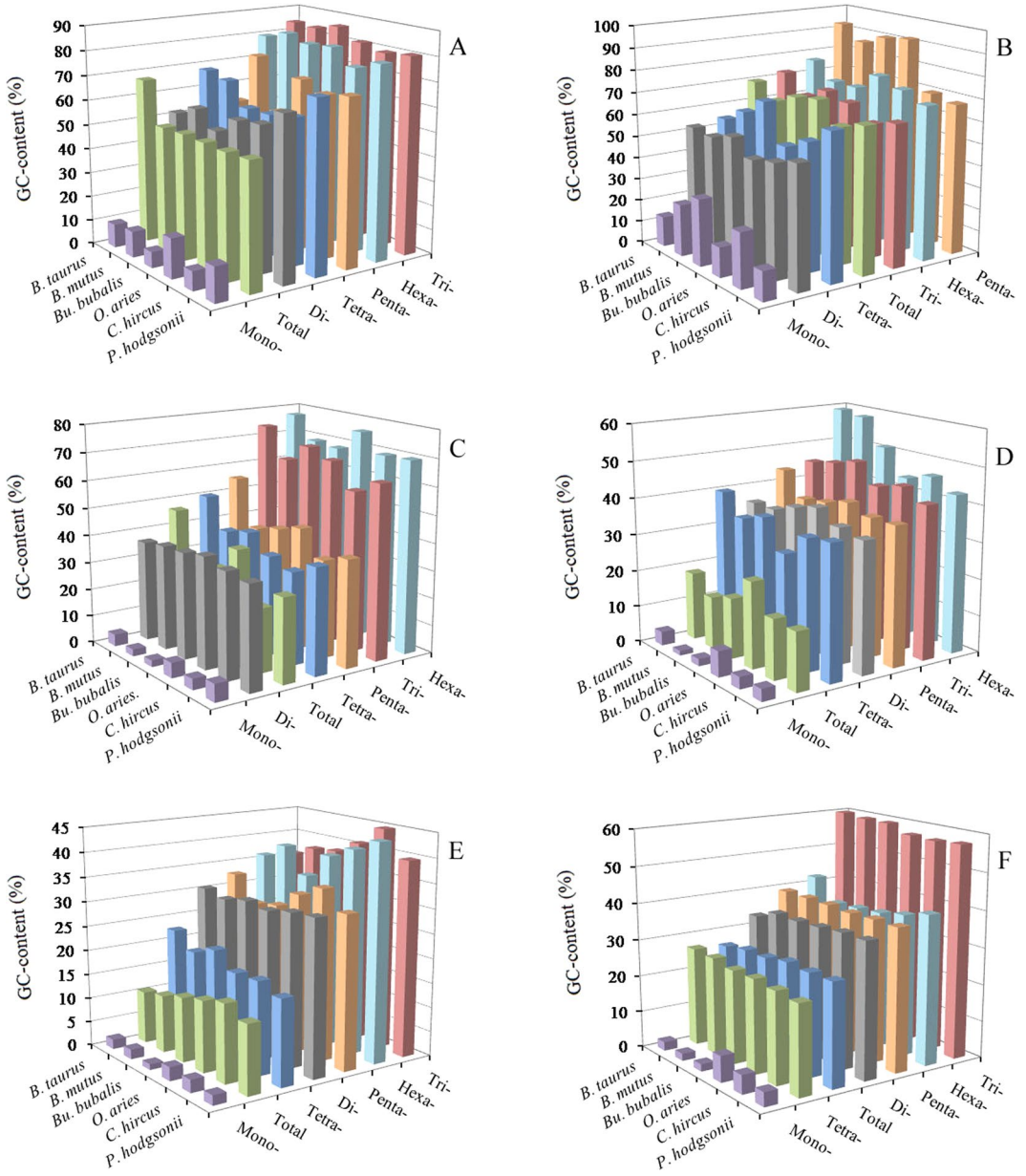

Figure 4. GC-contents of mono- to hexanucleotide P-SSRs in different intragenic and intergenic regions of six bovid species. ABCDEF represent 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions, respectively.

compared to other genomic regions in the bovid species. The $(CG)_n$ are relatively frequent in the 5′UTRs, whereas their abundance are very little in the coding regions, introns, 3′UTRs, TEs, and intergenic regions of the bovid species, this is consistent with the intragenic and intergenic regions of primates[22]. The $(A)_n$ repeat units are the most abundant motifs in the 5′UTRs, introns, 3′UTRs, TEs, and intergenic regons of these bovid species, this is consistent with bovid geomes[23]. The second most frequent motifs are dinucleotide $(AC)_n$ repeats in the introns, 3′UTRs, and intergenic regions of these bovid species, this is consistent with previous reports[22,23]. $(ACG)_n$ and $(AGC)_n$ motifs are comparatively frequent in intronic and intergenic regions of these bovid species, where their occurrence exceeds that of other trinucleotide repeat units. The $(CCG)_n$ motifs are the most abundant repeat units in 5′UTRs, the second in the coding regions; whereas the $(CCG)_n$ motifs are relatively infrequent in the introns, TEs, and intergenic regions, and also their abundance were less than that of other trinucleotide motifs in the bovid species. This is consistent with the different genomic regions of primates[22]. It has been demonstrated that the $(CCG)_n$ motif was significantly presented in the upstream regions of the genes[24]. The distributional pattern of SSR motifs in different genomic regions may be correlated with the present frequency of certain amino acids.

The variation of SSR abundance in different intragenic and intergenic regions. It has recently been reported that the distribution of SSRs is nonrandom in the genome, and their abundances vary widely in

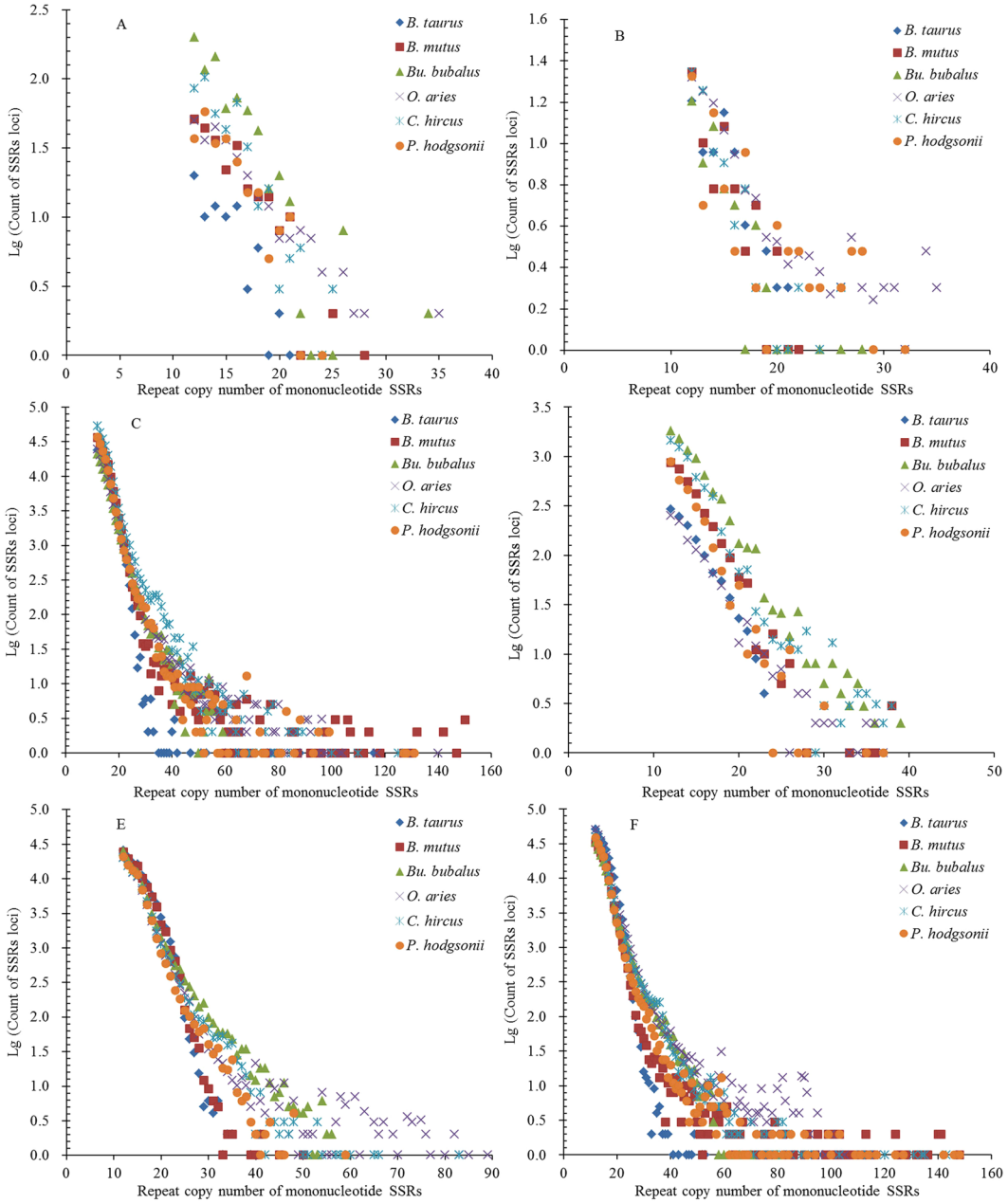

Figure 5. Comparative analysis of repeat copy number (RCN) of mononucleotide P-SSRs in different genomic regions of six bovid genomes. ABCDEF represent 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions, respectively.

different genomic regions[22]. Recent evidences have demonstrated that SSRs in different genomic regions play different functional roles. Many SSRs exist in ORFs of higher eukaryotes[21,25–27]. Consistent with previous studies in primates and plants[22,27], SSR abundance differs in 5′UTR and 3′ UTR regions of these bovid genomes. In the primates, trinucleotide SSRs show around double greater frequency in the 5′UTRs than that in the coding regions, whereas the latter had much more frequent trinucleotide P-SSRs than that in the intron, 3′UTRs, TEs[22]. Dominance of trinucleotide SSRs over other nucleotide units in coding regions may be caused by frameshift mutations to suppress non-trimeric SSRs in coding regions[28]. In *Arabidopsis thaliana*, low SSR abundances occurred in the centromeric region[29]. In *Drosophila melanogaster*, SSR distribution differs between X-chromosomes and autosomes[30]. Inconsistent with previous report[25,27], the distributions of SSRs showed great similarity in the intronic and intergenic sequences of these bovid genomes. These reports suggest a significant heterogeneity of SSR distribution in different genomic regions of organism genomes.

It has been reported that changes of SSRs are involved in several human diseases[31–33]. Our results showed that the abundance of different SSR motifs varies with the genomic regions. SSRs have been shown to be more abundant in non-coding regions than that in coding regions[21,25,27,34]. In the different genomic regions of the same bovid species, the introns, 3′UTRs, and intergenic regions had the most abundant P-SSRs, followed by the pattern:

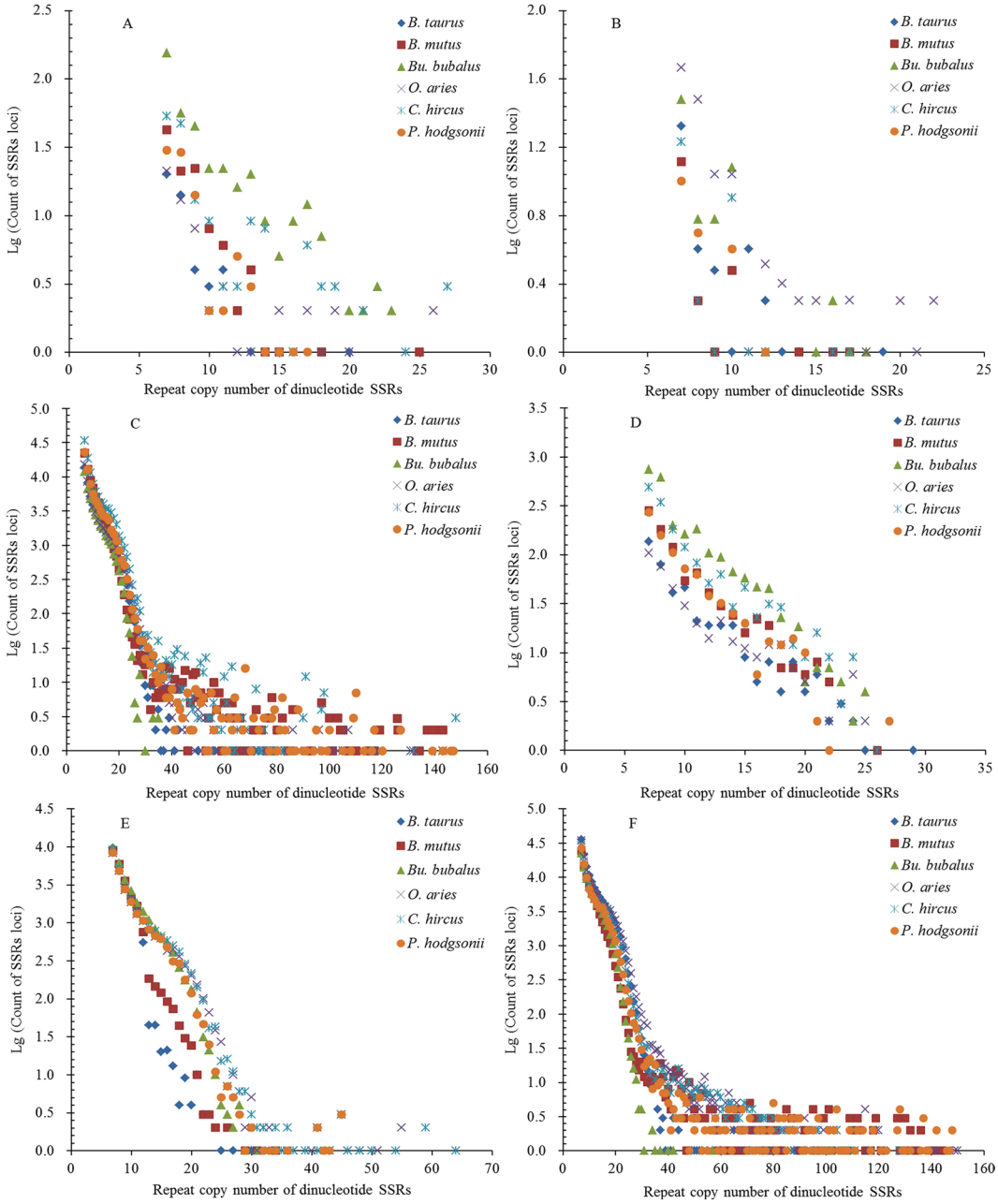

Figure 6. Comparative analysis of RCN of dinucleotide P-SSRs in different genomic regions of six bovid genomes. ABCDEF represent 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions, respectively.

5′UTRs > TEs > coding regions. There seem to be no distinct differences in P-SSR abundance between intronic and intergenic regions, which is consistent with previous report[25]. P-SSR abundance is the least in the coding regions, suggesting that low SSR abundance may decrease the evolvability of proteins. This may be related to the fact that SSR births/deaths were strongly selected against in coding regions[35].

This evidence has been proved that the mutations of coding regions could cause protein functional changes, loss of function, and protein truncation[4]. In different repeat type of these bovid species, trinucleotide P-SSRs were the most abundant type in the coding regions, whereas mononucleotide P-SSRs were the most frequent type in the 5′UTRs, introns, 3′UTRs, TEs, and intergenic regions; pentanucleotide P-SSRs were the least in the coding regions, whereas hexanucleotide P-SSRs were the least in the 5′UTRs, introns, 3′UTRs, TEs, and intergenic regions. In *Brassica rapa*, Trinucleotide SSRs were also the most frequent type in the coding regions[36]. In the exon regions, mononucleotide P-SSRs were the most abundant, followed by the pattern: tri- di- > tetra- > penta- > hexanucleotide SSRs in these bovid species. The abundances of hexanucleotide P-SSRs were less in the introns than that in the exons in these bovid species, which was inconsistent with previous reports[25]. It has been reported that coding regions are preferentially selected with trifold nucleotide SSR motifs[7,37–40] and suppressed non-trimeric

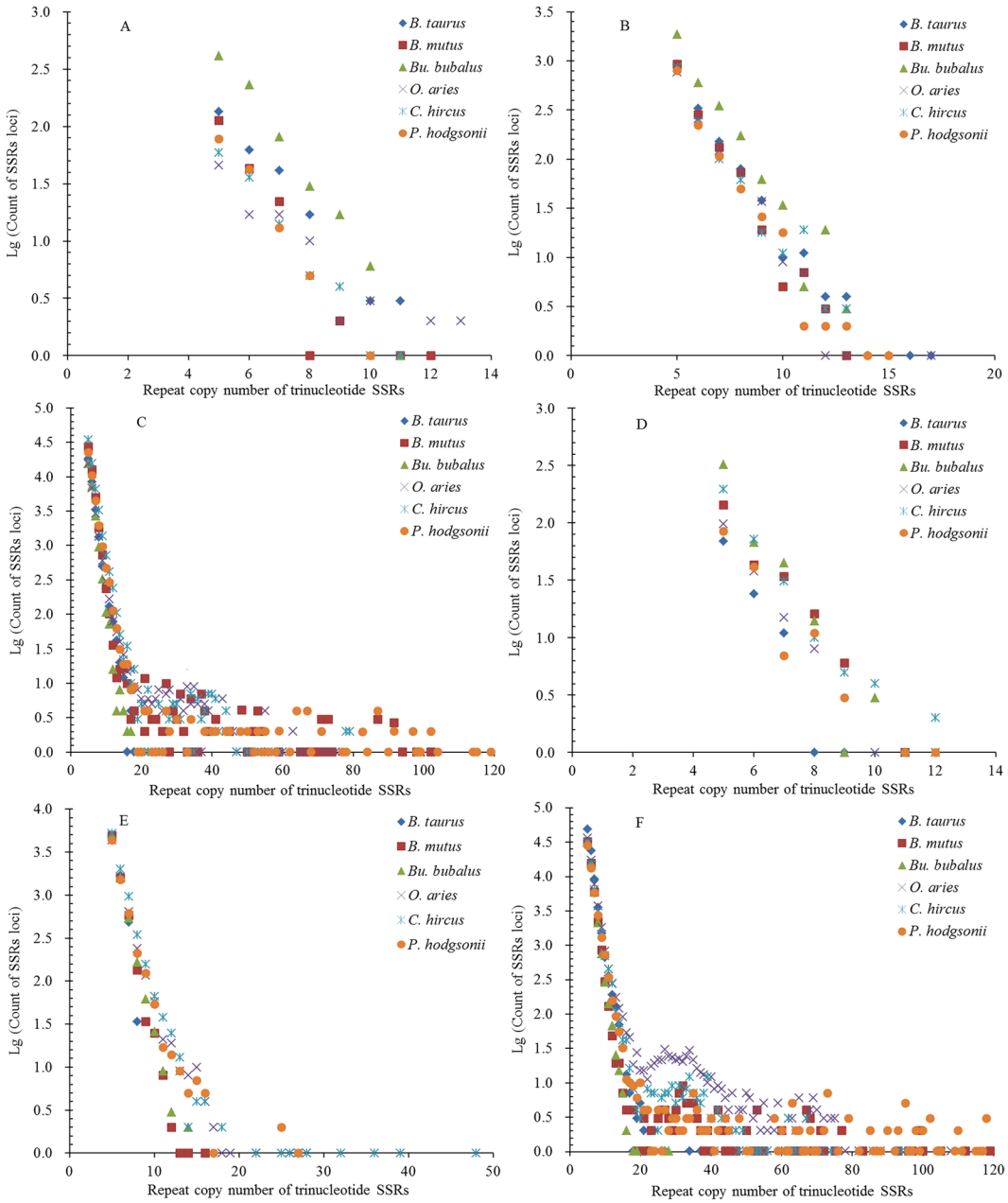

Figure 7. Comparative analysis of RCN of trinucleotide P-SSRs in different genomic regions of six bovid genomes. ABCDEF represent 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions, respectively.

SSR repeat units, which can reduce potential translational frameshift mutations[28]. This evidence can contribute to explain why trifold nucleotide SSR repeat units are more frequent in coding regions than that in other genomic regions.

The distributional pattern of GC-content in different genomic regions. Nucleotide composition influences SSR abundance, thus, the GC-content was examined in different genomic regions of six bovid species. The GC-contents of six bovid genomes showed to be remarkably consistent, but GC-contents varied greatly among different genomic regions. In this study, 5′UTRs had the most GC-content, followed by the coding regions (51.09–53.60%), thirdly the 3′UTRs and TEs, the least was the intronic and intergenic regions. Thus we can know that high GC-content was frequently distributed in exon-rich regions, and the distribution of GC-content was uneven in the bovid genomes. This evidence was consistent with the GC-content distributional pattern of different genomic regions in the primates[22]. Different classes of TEs tend to have bias for either GC-rich or GC-poor regions[41]. Ancestral Alu sequences have a high GC-content[42,43]. In the study, the repeat units of GC-richness were present in the 5′UTRs and coding regions, in which the GC-content were much higher than that in the remaining genomic regions (Fig. 4); whereas the motifs of AT-richness were present in the introns, 3′UTRs, TEs,

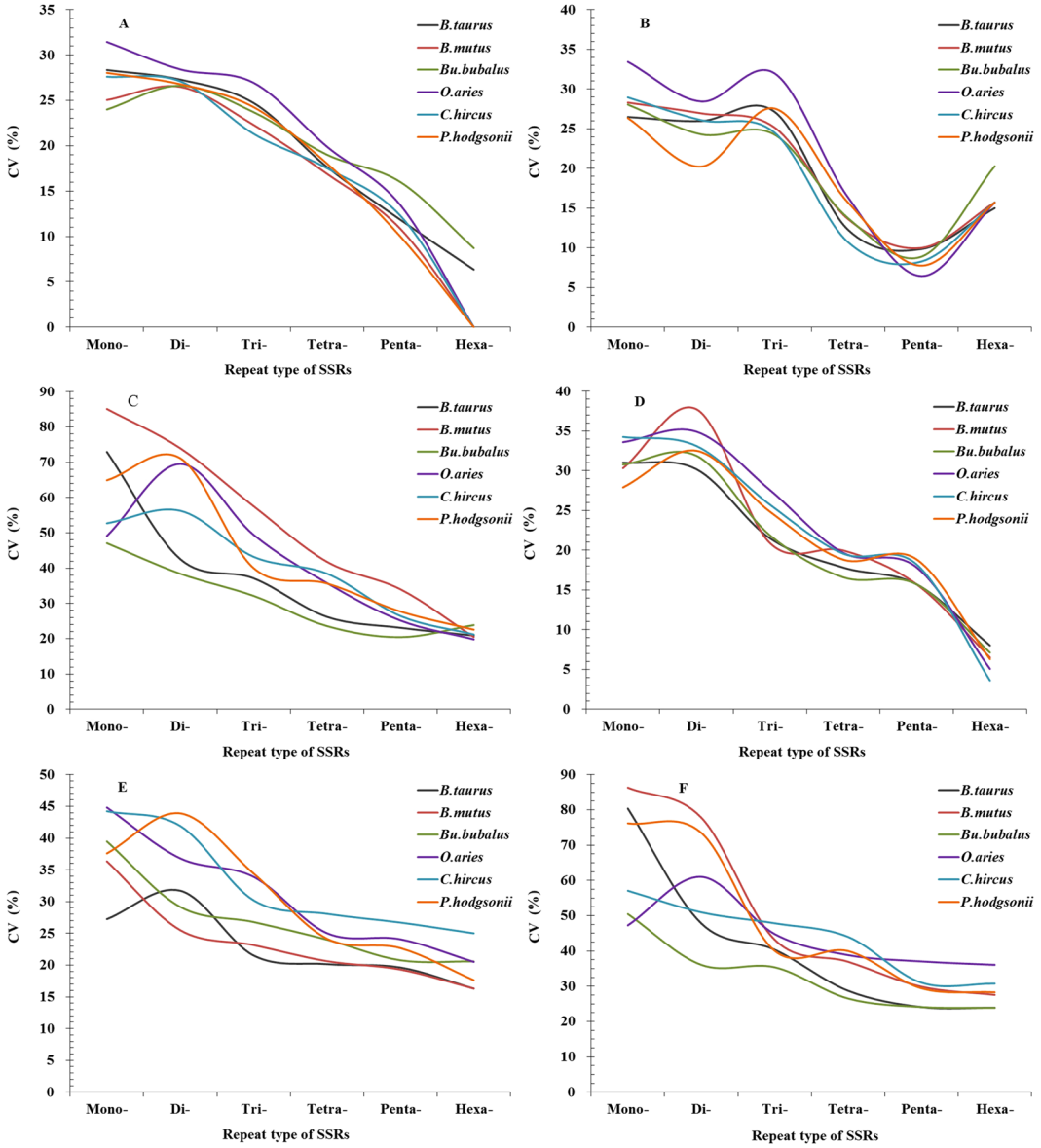

Figure 8. The CV analysis of RCN of SSRs in different genomic regions of six bovid genomes. ABCDEF represent 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions, respectively.

and intergenic regions, in which the AT-content were much higher than that in the 5′UTRs and coding regions (Tables S8–13). It has recently been reported that top SSR motifs have a direct positive relationship with the GC- or AT-content in different genomic regions[44]. In contrast, the gradient of average GC-content decreases from the 5′UTRs to intronic regions by several percent to around 14.88% in these different genomic regions of the bovids. It has been reported that there is a gradient in the GC-content of Gramineae genes[45]. It has also been reported that SSR polymorphism was negatively correlated with the GC-content of the flanking regions of SSR locus[46]. Furthermore, the GC-content of different genomic regions in the genome could be used as a relative measure of mutation rate.

Association of SSRs with other sequence elements and their mutability. SSRs associate characteristically with different intragenic and intergenic regions in the genome. SSR abundance is considerably high in 5′UTRs of plant genes[47,48] and are relatively low in exonic sequences[47]. SSRs are richly distributed in the 5′UTRs, introns, 3′UTRs, and intergenic regions of primates, and are relatively few in the coding regions[22]. SSR distribution in introns is similar to that of the whole genome[22,47]. Genomic regions of SSR collection have been recognized in *Arabidopsis thaliana*[49], *Drosophila melanogaster*[50], and primates[22]. In 42 prokaryotic genomes, SSR distributions in coding regions were biased toward coding termini[51]. SSRs are also frequently found in the proximity of TEs[52–54]. It has been confirmed that SSRs are often associated with retrotransposons[55], Alu elements, SINEs (short interspersed elements)[56], MITES (miniature inverted transposable elements)[47,55]. $(GAA)_n$ were associated with Alu repeats[56]. Abundant trinucleotide SSRs are distributed near genes[48,57], and tri- and hexanucleotide SSRs

predominated in the coding regions of these bovids. In the study, we have demonstrated that SSRs are obvious correlated with TEs (Fig. 1E and 2E).

The birth or death of SSRs is seemingly regulated by polymerase slippage, point mutations, and other activities involving chromatin reorganization[58,59]. SSR loci have a high mutation rate (10^{-6} to 10^{-2}/generation) which is due to strand slippage and unequal recombination leads to indels of repeat units[3]. The mutation rates associated with SSR loci are influenced by motif length, repeat number, and repeat type[60–63]. Mutation rates increase or decrease SSR repeat number, which are both frequent and reversible. Long SSR alleles have a downward mutation rates, which could result in a size constraint of SSRs[64–68]. Mutation rates also vary for different SSR loci within the same species[69]. There have been reported that a differential mutability rates for different SSRs occur in the genomes of two subspecies of rice[47]. Evolutionary dynamics of SSRs was regulated by their neighboring sequences[63]. SSR mutation rates vary obviously across the genomes. The abundance of tri- and hexanucleotide in coding regions also supported that specific selection against frameshift mutations in coding regions[4,22,28]. Trifold SSRs had not generated frameshifts through expansion of triplet SSRs, so that which would refrain from selective pressures in coding regions. However, non-trifold SSRs had to be subject to greater selection with the frameshift mutations[28]. RCN mutations of non-trifold SSRs in coding cause frameshifts, which can effectively inactivate gene expression or code for different or shorter protein sequences[1]. Therefore, mutation pressure contributed to the abundance of trifold SSRs in coding regions. SSR mutability per motif is relative higher at longer allele lengths[70]. Greater mutability per RCN was demonstrated in orthologous allele lengths between species[70]. These evidences have been demonstrated that SSR mutation process is great heterogeneous[70], showing differences in mutability between different allele lengths and motif sizes and between species.

Material and Methods

The sequences of intragenic and intergenic regions. We selected whole genome sequences of six bovids as subjects to analyze the SSR distribution of different genomic regions. The bovid genome sequences were downloaded in FASTA format from the Ensembl (http://asia.ensembl.org/index.html) and NCBI (https://www.ncbi.nlm.nih.gov/). The sequences of the gene models, 5′UTRs, coding regions, introns, 3′UTRs, TEs, and intergenic regions were generated according to the positions in the genome annotations. The intergenic regions referred to the interval sequences between gene and gene that were not comprised of the introns, coding regions, UTRs, and other related sequences. SSRs can be grouped into six categories[23,61,71], which were identified and scanned for SSRs of 1–6 bp using the software MSDB (Microsatellite Search and Building Database)[72] and Krait[16]. To compare our results, the same tool and search parameters were used in the data analysis of these bovid genomes.

SSRs identification and investigation. Since bovid species are large genomes, relatively systemic search criteria[72] were adopted in the study. In this study, repeat units with being circular permutations and/or reverse complements of each other were grouped together as one repeat unit for statistical analysis[73,74]. For tetra- and hexanucleotide repeat units, relatively systemic combination criteria were applied[23] in the process of filtration. For the sake of comparative analysis among different repeat types or motifs, relative abundance was determined, which means the number of SSRs per Mb of the sequence analyzed[72,75]. These total numbers have been normalized as relative abundance to allow comparison in the different genomic regions. In the four DNA bases, percentage of guanine (G) plus cytosine (C) was called GC-content in the analyzed sequence.

Variation analysis of SSRs. In order to analyze the variation of RCN of different repeat SSR types in the different genomic regions, we introduce the CV, which the calculation formula is as follow:

$$CV = S/\overline{x} \times 100\%.$$

where S is the standard deviation of the RCN of one SSR, $\overline{x}$ is the average of the RCN. The variation of RCN of two or more SSRs were comparative analyzed by the CV, which can eliminate the effect of different unit and mean, and is able to really reflect variation level of RCN of different SSRs.

References

1. Kashi, Y. & King, D. G. Simple sequence repeats as advantageous mutator in evolution. *Trend Genet.* **22**, 253–259 (2006).
2. Mrázek, J., Guo, X. & Shah, A. Simple sequence repeats in prokaryotic genomes. *P. Natl. Acad. Sci. USA* **104**, 8472–8477 (2007).
3. Deback, C. *et al*. Utilization of microsatellite polymorphism for differentiating herpes simplex virus type 1 strains. *J. Clin. Microbiol.* **47**, 533–540 (2009).
4. Li, Y. C., Korol, A. B., Fahima, T. & Nevo, E. Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol.* **21**, 991–1007 (2004).
5. Dresselhaus, T. *et al*. Novel ribosomal genes from maize arediferentiallyexpressedin thezygoticandsomaticcellcycles. *Mol Gen Genet.* **261**, 416–427 (1999).
6. Bao, J., Corke, H. & Sun, M. Microsatellites in starch-synthesizing genes in relation to starch physicochemical properties in waxy rice (*Oryza sativa* L.). *Theor. Appl. Genet.* **105**, 898–905 (2002).
7. Zhang, L. *et al*. Conservation of noncoding microsatellites in plants: implication for gene regulation. *BMC Genomics.* **7**, 323 (2006).
8. Perutz, M. F., Pope, B. J., Owen, D., Wanker, E. E. & Scherzinger, E. Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid β-peptide of amyloid plaques. *P. Natl. Acad. Sci. USA* **99**, 5596–5600 (2002).
9. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. *The FEBS journal* **272**, 5129–5148 (2005).
10. Timchenko, L. T. & Caskey, C. T. Trinucleotide repeat disorders in humans: discussions of mechanisms and medical issues. *FASEB J.* **10**, 1589–1597 (1996).
11. Sharma, P. C., Grover, A. & Kahl, G. Mining microsatellites in eukaryotic genomes. *Trends Biotechnol.* **25**, 490–498 (2007).
12. Thiel, T., Michalek, W., Varshney, R. K. & Graner, A. Exploiting est databases for the development and characterization of gene-derived ssr-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* **106**, 411–422 (2003).

13. Kofler, R., Schlotterer, C. & Lelley, T. SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics* **23**, 1683–1685 (2007).
14. Faircloth, B. C. Msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Mol. Ecol. Resour.* **8**, 92–94 (2008).
15. Wang, X. & Wang, L. GMATA: An Integrated software package for genome-scale SSR mining, marker development and viewing. *Front. Plant Sci.* **7**, 1350 (2016).
16. Du, L., Zhang, C., Liu, Q., Zhang, X. & Yue, B. Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics* **34**, 681–683 (2018).
17. Verkerk, A. J. *et al*. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell.* **65**, 905–914 (1991).
18. La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E. & Fischbeck, K. H. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* **352**, 77–79 (1991).
19. Zoghbi, H. Y. & Orr, H. T. Glutamine repeats and neurodegeneration. *Annu. Rev. Neurosci.* **23**, 217–247 (2000).
20. Sureshkumar, S. *et al*. A genetic defect caused by a triplet repeat expansion in *Arabidopsis thaliana*. *Science* **323**, 1060–1063 (2009).
21. Katti, M. V., Ranjekar, P. K. & Gupta, V. S. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol Biol Evol.* **18**, 1161–1167 (2001).
22. Qi, W. H. *et al*. Distinct patterns of simple sequence repeats and GC distribution in intragenic and intergenic regions of primate genomes. *Aging (Albany NY)* **8**, 2635–2650 (2016).
23. Qi, W. H. *et al*. Genome-Wide Survey and Analysis of Microsatellite Sequences in Bovid Species. *Plos One.* **10**, e0133667 (2015).
24. Subramanian., S. *et al*. Triplet repeats in human genome: distribution and their association with genes and other genomic regions. *Bioinformatics* **19**, 549–552 (2003).
25. Tóth, G., Gáspári, Z. & Jurka, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10**, 967–981 (2000).
26. Kantety, R. V., La Rota, M., Matthews, D. E. & Sorrells, M. E. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol. Biol.* **48**, 501–510 (2002).
27. Morgante, M., Hanafey, M. & Powell, W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **30**, 194–200 (2002).
28. Metzgar, D., Bytof, J. & Wills, C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* **10**, 72–80 (2000).
29. Schlötterer, C. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**, 365–371 (2000).
30. Bachtrog, D., Agis, M., Imhof, M. & Schlötterer, C. Microsatellite variability differs between dinucleotide repeat motifs-evidence from Drosophila melanogaster. *Mol. Biol. Evol.* **17**, 1277–1285 (2000).
31. Utsch, B. *et al*. A novel stable polyalanine [poly(A)] expansion in the HOXA13 gene associated with hand-foot-genital syndrome: proper function of poly(A)-harbouring transcription factors depends on a critical repeat length? *Hum. Genet.* **110**, 488–494 (2002).
32. Hancock, J. M. & Simon, M. Simple sequence repeats in proteins and their significance for network evolution. *Gene* **345**, 113–118 (2005).
33. Pearson, C. E., Edamura, K. N. & Cleary, J. D. Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.* **6**, 729–742 (2005).
34. Lawson, M. J. & Zhang, L. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biol.* **7**, R14 (2006).
35. Kelkar, Y. D., Eckert, K. A., Chiaromonte, F. & Makova, K. D. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Res.* **21**, 2038–2048 (2011).
36. Hong, C. P. *et al*. Genomic distribution of simple sequence repeats in *Brassica rapa*. *Mol. Cells* **23**, 349–56 (2007).
37. Fujimori, S. *et al*. A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. *Febs Lett.* **554**, 17–22 (2003).
38. Subramanian, S., Mishra, R. K. & Singh, L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol.* **4**, R13 (2003).
39. Li, B. Q., Xia, C., Lu, Z., Zhou, Z. & Xiang, Z. Analysis on frequency and density of microsatellites in coding sequences of several eukaryotic genomes. *Genomics, Proteomics & Bioinformatics* **2**, 24–31 (2004).
40. Mayer, C., Leese, F. & Tollrian, R. Genome-wide analysis of tandem repeats in Daphnia pulex-a comparative approach. *BMC Genomics* **11**, 277 (2010).
41. Arndt, P. F., Hwa, T. & Petrov, D. A. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J. Mol. Evol.* **60**, 748–763 (2005).
42. Britten, R. J., Baron, W. F., Stout, D. B. & Davidson, E. H. Sources and evolution of human Alu repeated sequences. *P. Natl. Acad. Sci. USA* **85**, 4770–4774 (1988).
43. Jurka, J. & Smith, T. A fundamental division in the Alu family of repeated sequences. *P. Natl. Acad. Sci. USA* **85**, 4775–4778 (1988).
44. Victoria, F. C., da Maia, L. C. & de Oliveira, A. C. *In silico* comparative analysis of SSR markers in plants. *BMC Plant Biol.* **11**, 15 (2011).
45. Wong, G. K. S. *et al*. Compositional gradients in Gramineae genes. *Genome Res.* **12**, 851–856 (2002).
46. Glenn, T. C., Stephan, W., Dessauer, H. C. & Braun, M. J. Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. *Mol. Biol. Evol.* **13**, 1151–1154 (1996).
47. Grover, A., Aishwarya, V. & Sharma, P. C. Biased distribution of microsatellite motifs in the rice genome. *Mol. Genet. Genomics* **277**, 469–480 (2007).
48. Zhang, L. *et al*. Preference of simple sequence repeats in coding and non-coding regions of Arabidopsis thaliana. *Bioinformatics* **20**, 1081–1086 (2004).
49. Grover, A. & Sharma, P. C. Microsatellite motifs with moderate GC content are clustered around genes on Arabidopsis thaliana chromosome 2. *In Silico Biology* **7**, 201–213 (2007).
50. Bachtrog, D., Weissm, S., Zangerl, B., Brem, G. & Schlötterer, C. Distribution of dinucleotide microsatellites in the Drosophila melanogaster genome. *Mol. Biol. Evol.* **16**, 602–610 (1999).
51. Lin, W. H. & Kussell, E. Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. *Nucleic. Acids Res.* **40**, 2399–2413 (2012).
52. Arcot, S. S., Wang, Z., Weber, J. L., Deininger, P. L. & Batzer, M. A. Alu repeats: a source for the genesis of primate microsatellites. *Genomics* **29**, 136–144 (1995).
53. Ramsay, L. *et al*. Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. *Plant J.* **17**, 415–425 (1999).
54. Temnykh, S. *et al*. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.* **1**, 1441–1452 (2001).
55. Schulman, A. H., Gupta, P. K. & Varshney. R. K. Organization of retrotransposons and microsatellites in cereal genomes (ed. Gupta, P. K. & Varshney, R. K. Cereal Genomics) 83–118 (Springer, 2005).
56. Wang, H. *et al*. SVA elements: a hominid-specific retroposon family. *J. Mol. Biol.* **354**, 994–1007 (2005).

57. Jayashree, B. *et al*. A database of simple sequence repeats from cereal and legume expressed sequence tags mined *in silico*: survey and evaluation. *In Silico. Biology* **6**, 607–620 (2006).
58. Calabrese, P. P., Durrett, R. T. & Aquadro, C. F. Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. *Genet.* **159**, 839–852 (2001).
59. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *P. Natl. Acad. Sci. USA* **101**, 12404–12410 (2004).
60. Brown, L. Y. & Brown, S. A. Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends Genet.* **20**, 51–58 (2004).
61. Chambers, G. K. & MacAvoy, E. S. Microsatellites: consensus and controversy. *Comp. Biochem. Physiol. B.* **126**, 455–476 (2000).
62. Vergnaud, G. & Denoeud, F. Minisatellites: Mutability and genome architecture. *Genome Res.* **10**, 899–907 (2000).
63. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5**, 435–445 (2004).
64. Schlötterer, C. Are microsatellites really simple sequences? *Curr.Biol.* **8**, R132–R134 (1998).
65. Ellegren, H. Heterogeneous mutation processes in humanmicrosatellite DNA sequences. *Nat. Genet.* **24**, 400–402 (2000).
66. Harr, B. & Schlötterer, C. Long microsatellite alleles inDrosophila melanogaster have a downward mutation bias and shortpersistence times, which cause their genome-wideunderrepresentation. *Genetics* **155**, 1213–1220 (2000).
67. Xu, X., Peng, M. & Fang, Z. The direction of microsatellitemutations is dependent upon allele length. *Nat. Genet.* **24**, 396–399 (2000).
68. Harr, B., Todorova, J. & Schlötterer, C. Mismatch repair drivenmutational bias in D. melanogaster. *Mol. Cell.* **10**, 199–205 (2002).
69. Whittaker, J. C. *et al*. Likelihood-based estimation of microsatellite mutation rates. *Genetics* **164**, 781–787 (2003).
70. Webster, M. T., Smith, N. G. C. & Ellegren, H. Microsatellite evolution inferred from human–chimpanzee genomic sequence alignments. *P. Natl. Acad. Sci. USA* **99**, 8748–8753 (2002).
71. Bachmann, L., Bareiss, P. & Tomiuk, J. Allelic variation, fragment length analyses and population genetic models: a case study on *Drosophila* microsatellites. *J. Zool. Syst. Evol. Res.* **42**, 215–223 (2004).
72. Du, L., Li, Y., Zhang, X. & Yue, B. MSDB: A user-friendly program for reporting distribution and building databases of microsatellites from genome sequences. *J. Hered.* **104**, 154–157 (2013).
73. Jurka, J. & Pethiyagoda, C. Simple repetitive DNA sequences from primates: compilation and analysis. *J. Mol. Evol.* **40**, 120–126 (1995).
74. Li, C. Y. *et al*. Genome-wide analysis of microsatellite sequence in seven filamentous fungi. *Interdiscip Sci.* **1**, 141–150 (2009).
75. Karaoglu, H., Lee, C. M. Y. & Meyer, W. Survey of simple sequence repeats in completed fungal genomes. *Mol. Biol. Evol.* **22**, 639–649 (2005).

Acknowledgements

The research supported by the Foundation of Key Laboratory of Southwest China Wildlife Rsources Conservation (Ministry of Education, XNYB16-5) and National Natural Science Foundation (NSFC31702032) of P. R. China. Furthermore, this work was funded by the Scientific and Technological Research Program of Chongqing Municipal Education Commission (No.KJ1710239), Major breeding Project (No. 15ZP03) of Chongqing Three Gorges University, Basic and Advanced Research Project of Chongqing science &technology Commission (No. cstc2015jcyjA80016), P. R. China. We thank Charles Wayne Wilson at Chongqing Three Gorges University, Thomas Connor at Michigan State University, and Timothy Moermond at Sichuan University, for revising our manuscript.

Author Contributions

Conceived and designed the experiments: W.H.Q., C.Q.Z. and B.S.Y. Performed the experiments: W.H.Q. and C.C.Y. Analyzed the data: X.M.J. and W.H.Q. Contributed reagents/materials/analysis tools: C.C.Y. Wrote the paper: W.H.Q. Completed all figures: W.Q.Z. Collected Bovidae genomes and performed the SSRs validation: G.S.X., W.Q.Z.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-32286-5.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.