Genome-wide mining and comparative analysis of microsatellites in five domestic birds

Xing Yue Li[1, *], Hao Xiang Wang[1] and Lan Ou[2]

[1]School of Life Sciences, Sichuan University, Chengdu, China

[2]School of Computer Science, Sichuan University, Chengdu, China

*Corresponding author: hurchali@163.com

Abstract. Microsatellites are widely present in the gene sequences of most species and are closely associated with genome structure, function and certain diseases. To identify microsatellites in five domesticated bird species, we searched and compared Amazona aestiva, Aaes; Myiopsitta monachus, Mmon; Acridotheres tristis, Atri; Anas platyrhynchos, Apla; Gallus gallus. Perfect microsatellites (P-SSR) with nucleotide lengths between 1 and 6 in the Ggal gene sequence were identified. A total of 101,395, 171,161, 226,953, 458,231 and 272,487 P-SSRs were identified. the most common perfect SSRs were single nucleotide SSRs. the next most common SSRs for Aaes, Mmon and Apla were tetranucleotides (11.30%, 15.22% and 15.62%) the least while for Atri and Ggal they were pentanucleotides (23.67% and 19.57%) and the least were dinucleotide SSRs (5.75%) and trinucleotide SSRs (5.89%), respectively. In the GC content statistics for all SSRs, the total GC content accounted for 1.95% (Aaes), 3.19% (Mmon) 29.71% (Atri), 16.73% (Apla) and 27.39% (Ggal) of the total length of SSRs, respectively. The GC content of almost all types of SSRs was higher than that of the whole genome, except for single nucleotide SSRs, and the lowest GC content was found in single nucleotide SSRs. and, the single nucleotide GC content of Ggal (11.89%) was 45 times higher than that of Aaes (0.26%). The comparison of several aspects of the distribution frequency of SSRs, GC content and repeat sequences of nucleotides showed that the microsatellite distribution characteristics in the five species differed significantly between species. The above results and data reveal differences among the five bird species at the genomic level and genetic diversity of species in domesticated birds. Our data will contribute to comparative genome mapping, understanding the distribution and genomic structure of SSRs among these animal models, and laying the foundation for further development and identification of more domesticated bird-specific SSRs.

Keywords: Microsatellite; Nucleotides; Genome-wide mining.

1. Introduction

Microsatellites, also known as simple sequence repeats (SSR) or short tandem repeats (STR), have nucleotide motifs of 1-6 bp of DNA sequence and are widely distributed in eukaryotic and prokaryotic genomes (Oliveira et al. 2006). They are characterized by high frequency, good coordination and high polymorphism (Schlötterer 1998). Microsatellite expansions may be caused by point mutations (Meissier et al. 1996) and replication slips (Ellegren 2004; Choudhary and Trivedi 2010) and occur extremely slowly on evolutionary time scales. Since the invention of polymerase chain reaction (PCR) technology, microsatellites have been widely used in population genetics (Sunnucks 2000), individual recognition and parentage identification as genetic markers (Schlötterer 2000). SSRs are found not only in coding regions but also in non-coding regions (Li et al. 2004). Many studies have shown that SSRs located in promoter regions affect gene expression and that gene transcription is also affected by SSRs in introns (Chistiakov et al. 2006) [1].

In this paper, five common domestic birds were selected and their genomic microsatellite analysis was performed.

In this paper, we have done a systematic comparative study of genome-wide SSRs of five bird species. We were curious if the five domesticated bird species are similar to each other at the level of genomic SSRs or are there any significant differences?

In the past few years, several researchers have developed and characterized microsatellite markers for Anodorhynchus hyacinthinus (da Silva, Presti et al. 2015), providing valuable data support and a framework for research methods for SSR characterization of bird genomes, but so far, no genome-wide SSRs have been conducted for different birds that are adept at mimicking sounds Genome-wide identification and comparison of SSRs for five bird species were carried out specifically. Previous research explorations have focused on the specific use of microsatellites in developing molecular markers, such as wheat (Mehta, Muthusamy et al. 2021), Staphylinidae (Park, Son et al. 2019), and partially comparing SSRs in the genomes of different species, such as fish (Lei, Zhou et al. 2021), the Orf virus (Sahu, Majee et al. 2020), and macaques (Liu, Hou et al. 2017), while SSRs for birds have focused on various rare, endangered species (Jan and Fumagalli 2016), with studies suggesting that our knowledge of the genomes of domesticated birds is still very sparse (van der Zwan and van der Sluis 2021). This study provides strong support for our understanding of the gene distribution patterns of popular domesticated birds and further development of more and more effective molecular markers [2].

2. Materials and Methods

2.1. Sources of genomic data

Whole genome sequences of five bird species (Amazona aestiva, Aaes; Myiopsitta monachus, Mmon, Acridotheres tristis, Atri, Anas platyrhynchos, Apla, Gallus gallus, Ggal) were obtained from the NCBI online website for download (https://www.ncbi.nlm.nih.gov/). The assembly file numbers are GCA_017639355.1 (Aaes), GCA_017639245.1 (Mmon), GCA_027559615.1 (Atri), NC_051774.1 (Apla), and NW_024095933.1 (Ggal)

2.2. Identification and statistical analysis of microsatellite loci

The software used for searching and identifying microsatellite loci is MSDBv2.4.3. There are two modes of searching microsatellite loci, Perfect Search and Imperfect Search, and we first used Imperfect Search to search for six different types of loci, including perfect SSRs, interrupted perfect SSRs, compound SSRs, interrupted compound SSRs, complex SSRs, and interrupted complex SSRs, We further searched for microsatellites with different nucleotide repeat types using the Perfect Search mode. In terms of parameter settings, we set the minimum number of repeats for single to hexanucleotide SSRs to 12, 7, 5, 4, 4 and 4, respectively, based on previous studies, and the flanking sequence length to the software default value of 200 bp, and all other software parameters were set to default values. We used a python script to preprocess the raw data of microsatellites to count the circularly aligned sequences as well as their complementary sequences in the same SSR. For example, the cyclic sequence of ACG contains ACG, CGA and GAC, and their complementary sequences contain TGC, GCT and CTG. The above six sequences are considered to be the same microsatellite sequence after preprocessing the data. The run mode we set in the software will count the perfect microsatellite loci on all chromosomes. We also calculated the relative frequency and relative density of microsatellites for each chromosome and compared the microsatellite related information between the two species [3].

3. Results

3.1. The number, relative frequency, relative density and GC content of microsatellites in the three genomes

Microsatellite analysis was performed on the genomes of five bird species using the software MSDB 2.4.3. Among them, the most common types were all P-SSRs. for Aaes and Mmon ranked second were IP-SSRs and third were CD-SSRs, while the second most frequent in Apla and Ggal were CD-SSRs and the third most frequent were ICD-SSRs. while for Atri ranked second were ICD-

SSRs and third were IP-SSRs. and among them The least among them all are CX-SSRs (Table 1). The number and frequency of different microsatellites varied among different species, but the distribution of microsatellites was more similar in birds with closer relatives. The number, length, relative frequency and density of perfect microsatellites are shown in Table 1. A total of 101,395, 171,161, 226,953, 458,231 and 272,487 P-SSRs were identified with total frequencies of 2392.36 loci/Mb, 3726.92 loci/Mb, 209.93 loci/Mb, 385.54 loci/Mb and 258.69 loci/Mb, respectively. It can be seen that the number of SSRs of both parrots was less than other birds, but the frequencies were much higher than other species of birds.

Table 1. Number, length, relative frequency and density of perfect microsatellites

	CD-SSRs	CX-SSRs	ICD-SSRs	ICX-SSRs	IP-SSRs	P-SSRs
Aaes						
No. of SSRs	1,001	52	929	196	1,366	101,395
Frequency (loci/Mb)	19.09	1.75	19.5	5.07	28.95	2,392.36
Mmon						
No. of SSRs	2,073	180	1,979	740	2,099	171,161
Frequency (loci/Mb)	46.61	4.73	48.96	28.69	49.94	3,726.92
Atri						
No. of SSRs	3,867	381	6,079	3,403	4,880	226,953
Frequency (loci/Mb)	3.58	0.35	5.62	3.15	4.51	209.93
Apla						
No. of SSRs	12,784	760	9,583	3,848	7,522	458,231
Frequency (loci/Mb)	10.76	0.64	8.06	3.24	6.33	385.54
Ggal						
No. of SSRs	6,091	419	4,184	1,600	2,743	272,487
Frequency (loci/Mb)	5.78	0.4	3.97	1.52	2.6	258.69

From Figure 1, it can be seen that for these five bird species, the most common perfect SSRs were all mononucleotide SSRs and the highest percentage of them were for two parrots, accounting for 65.83% and 55.75% of the total SSRs. The second most common SSRs for Aaes, Mmon and Apla were tetranucleotide (11.30%, 15.22% and 15.62%) the least were hexanucleotide SSRs (1.58%, 2.47% and 4.83%). While for Atri and Ggal the second most common were pentanucleotide (23.67% and 19.57%), the least were dinucleotide SSRs (5.75%) and trinucleotide SSRs (5.89%), respectively. Their SSR numbers did not conform to a significant positive correlation with both relative frequency and density and varied significantly between species.

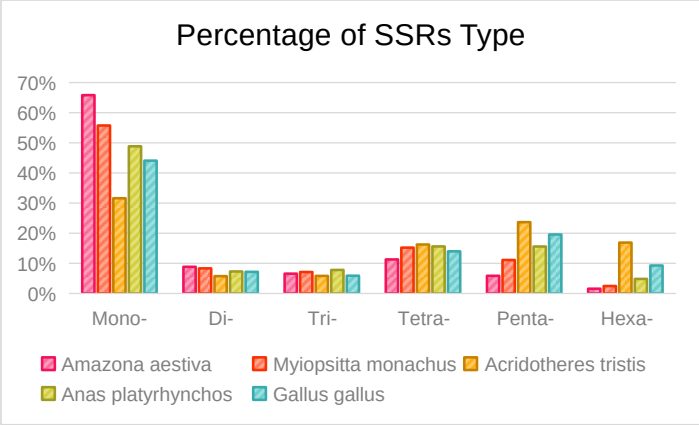

Fig. 1 Percentage of SSRs Type

We calculated the GC content of single to hexanucleotide perfect SSRs in both genomes (Table 2). The total GC content was 1.95% (Aaes), 3.19% (Mmon) 29.71% (Atri), 16.73% (Apla) and 27.39% (Ggal) of the total length of SSRs, respectively. Almost all types of SSRs had lower GC content than the whole genome and the lowest GC content was found in single nucleotide SSRs. single nucleotide GC content of Ggal (11.89%) was 45 times higher than that of Aaes (0.26%). In these genomes, CG

content also varied greatly among species in different repeat types, with Ggal reaching 48.51% and 46.91% CG content in pentanucleotide repeat and hexanucleotide repeat microsatellites.

Table 2. GC content of single nucleotide to hexanucleotide perfect SSRs

Repeat type		Total numbers	Total length (bp)	Frequency (loci/Mb)	Density (bp/Mb)	(G+C)	
						Length (bp)	Contenta (%)
Mono-	*Aaes*	76,558	1,249,443	1,543.91	24,509.043	3,317	0.26
	MMon	126,450	1,907,327	2,563.52	37,786.38	24,478	1.28
	Atri	139,538	2,189,356	129.07	2,025.11	96,661	4.42
	Alpa	326,911	5,497,171	275.05	4,625.172	283,525	5.16
	Ggal	192,334	3,115,540	182.6	2,957.794	370,286	11.89
Di-	*Aaes*	9,726	174,216	289.57	5,139.014	5,281	3.03
	MMon	15,843	281,632	312.13	5,476.87	9,850	3.50
	Atri	21,332	398,410	19.73	368.521	114,729	28.8
	Alpa	43,618	822,056	36.7	691.656	174,241	21.2
	Ggal	26,910	507,466	25.55	481.772	144,339	28.44
Tri-	*Aaes*	7,543	132,417	313.74	5,535.301	9,092	6.87
	MMon	14,140	242,538	435.38	7,621.866	18,806	7.75
	Atri	20,128	403,821	18.62	373.526	163,622	40.52
	Alpa	41,637	878,850	35.03	739.441	235,837	26.83
	Ggal	23,065	416,301	21.9	395.223	148,926	35.77
Tetra-	*Aaes*	10,502	220,788	283.65	6,062.134	11,849	5.37
	MMon	21,242	521,764	560.95	13,503.99	30,733	5.90
	Atri	33,002	1,125,500	30.53	1,041.065	441,652	39.24
	Alpa	73,828	1,757,416	62.12	1,478.643	390,939	22.25
	Ggal	38,809	988,548	36.84	938.496	293,447	29.68
Penta-	*Aaes*	3,662	116,075	90.99	2,875.774	6,015	5.18
	MMon	7,425	374,380	174.73	7,244.391	17,261	4.61
	Atri	29,550	1,639,695	27.33	1,516.685	699,466	42.66
	Alpa	36,176	1,754,590	30.44	1,476.265	572,301	32.62
	Ggal	20,045	1,383,115	19.03	1,313.085	670,901	48.51
Hexa-	*Aaes*	842	32,310	23.78	856.615	1,937	5.99
	MMon	1,981	85,860	117.13	4,467.464	6,595	7.68
	Atri	28,044	1,169,958	25.94	1,082.188	699,466	46.34
	Alpa	12,138	543,336	10.21	457.148	572,301	41.48
	Ggal	5,481	656,304	5.2	623.074	670,901	46.91
Total	*Aaes*	108,833	1,925,249	2,545.64	44,977.881	37,491	1.95
	MMon	187,081	3,413,501	4,163.84	76,100.96	108,723	3.19
	Atri	271,594	6,926,740	251.22	6,407.095	542,150	29.71
	Alpa	534,308	11,253,419	449.55	9,468.325	225,360	16.73
	Ggal	306,644	7,067,274	291.12	6,709.444	307,895	27.39

3.2. The number, relative frequency, relative density and GC content of microsatellites in the three genomes

3.2.1 Single nucleotide repeat type microsatellite loci

As shown in Figure 2, among the single nucleotide repeats, the greater weight was given to T(n) repeats, which were repeated 66,631, 49,978, 15,4570, 85,631 and 51,776 times in Atri, Aaes, Apla, Ggal and MMon, respectively, accounting for 47.75%, 47.71%, 47.28%, 47.28% and 40.23% of the total single nucleotide SSRs of the five species, respectively , 44.52% and 40.23%, respectively. And the G(n) repeat type with a weight of 4.33% was the lowest repeat frequency among the genomes of these five species. Nevertheless, the frequency of G(n) repeat categories in MMon was nearly six times higher relative to the number in Aaes. The maximum number of single nucleotide repeat types was 215, 90, 444, 662 and 239 in Atri, Aaes, Apla, Ggal and MMon, respectively.

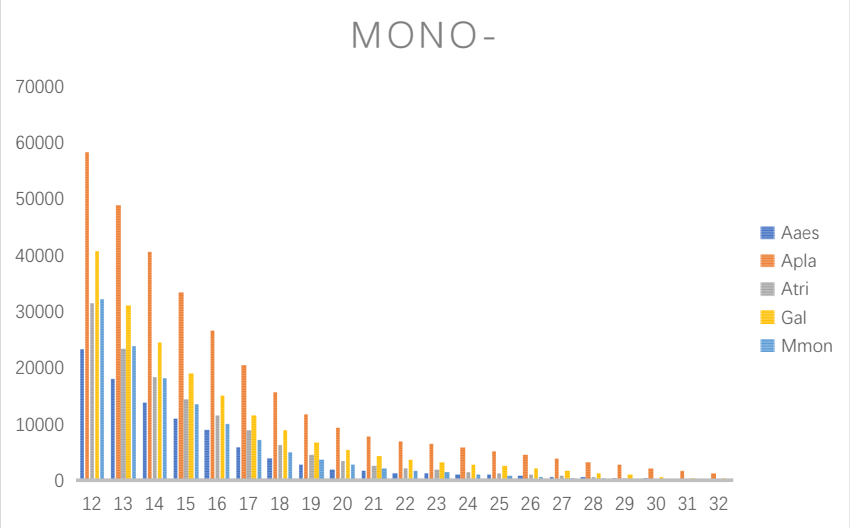

Fig. 2 Single nucleotide repeat type microsatellite loci (repeats>32 partially obscure and omitted)

3.2.2 Dinucleotide repeat type microsatellite loci

As shown in Figure 3, in Atri, AT (n) with a frequency of 4.58 loci/Mb was the most common dinucleotide SSR accounting for 23.17%, followed by TA (n), and CG (n) with the lowest frequency, nearly absent. In Aaes, the most common dinucleotide was TA (n) with 20.68% frequency of 2.41 loci/Mb, followed by TG (n), and the least frequent was CG (n). In Apla, the most common dinucleotide was AT (n) with 30.13% frequency of 11.05 loci/Mb, followed by TA (n) and the least by CG (n). In Ggal, the most common dinucleotide was TA (n) with 20.19% frequency of 5.16 loci/Mb, followed by AC (n) and the least by CG (n). In Mmon, the most common dinucleotide was AT(n) with 20.89% frequency of 3.01 loci/Mb, followed by TA(n) and the least by CG(n).

AT(n) repeats were the dominant dinucleotide SSR, accounting for 27.94 loci/Mb of dinucleotide sequences. the second most common was the TA(n) repeat sequence type with 22.89 loci/Mb, while TG(n) repeats occurred slightly less frequently than AT(n) and TA(n). CG(n) repeats were the least frequent of the five bird species sequences, accounting for 0.05loci/Mb of the dinucleotide sequences.

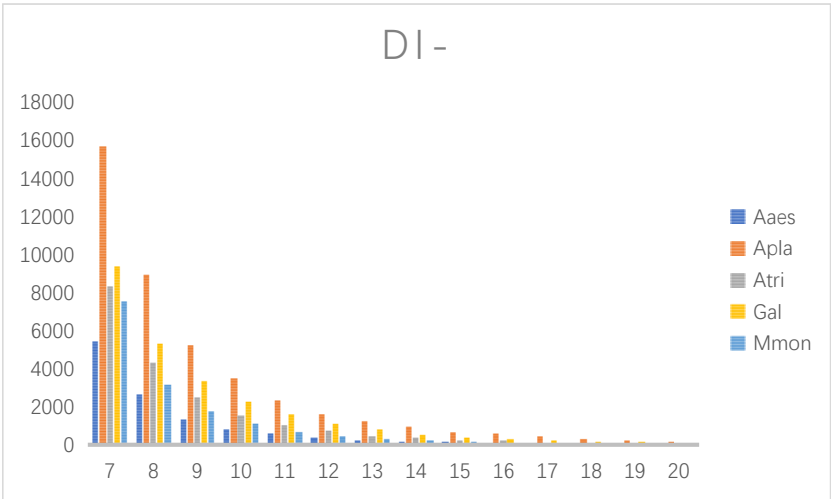

Fig. 3 Dinucleotide repeat type microsatellite loci (repeats>20 partially obscure and omitted)

3.2.3 Trinucleotide repeat type microsatellite loci

As shown in Figure 4, the most common trinucleotide in Atri is (TAA)n, with a density of 0.91loci/Mb. The most common trinucleotide in Aaes, Apla and Mmon is (ATT)n, with densities of 0.68loci/Mb, 3.93loci/Mb and 0.99loci/Mb, respectively, accounting for 7.79%, 11.23% and 7.84% of the total trinucleotide repeats in Aaes, Apla and Mmon, respectively. The most common

trinucleotide in Ggal was (GTT)n with a density of 2.14loci/Mb. The maximum number of trinucleotide repeats was 87, 25, 56, 36 and 37 in Atri, Aaes, Apla, Ggal and MMon, respectively.

We identified a total of six trinucleotide sequence repeats with density greater than 2 loci/Mb in Apla, namely (AAT)n, (ATT)n, (CAA)n, (GTT)n, (TAA)n, and (TTA)n. The (GTT)n repeat type also had the greatest frequency in Ggal with a density of more than 2.14 loci/Mb [5].

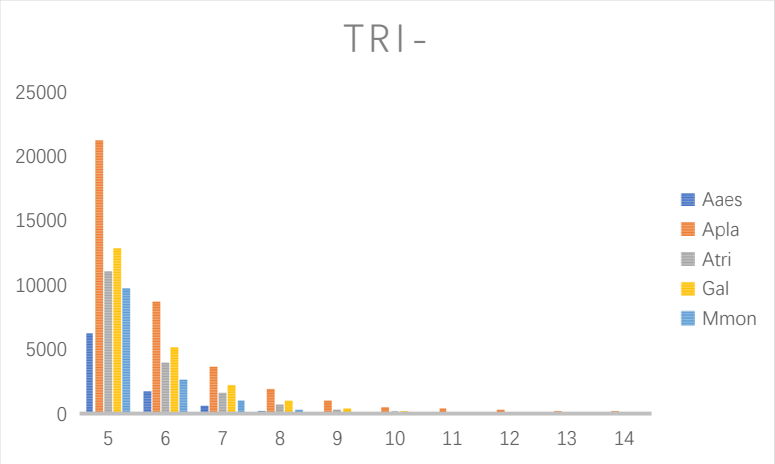

Fig. 4 Trinucleotide repeat type microsatellite loci (repeats>14 partially obscure and omitted)

3.2.4 Tetranucleotide repeat type microsatellite loci

As shown in Figure 5, for Aaes, Mmon, Apla and Ggal, the (AAAC)n repeats with frequencies of 1.9, 1.47, 8.38 and 5.07 loci/Mb were the most common tetranucleotide SSRs. but in Atri it was CATC (2.82 loci/Mb) In these five bird species, the highest frequency tetranucleotide repeats were 1157 (Mmon) and the second highest was 622 (Ggal).

Moreover, we found that for tetranucleotide repeats, we could not find any repeat with a frequency greater than 2loci/Mb in Aaes and Mmon, but two in Atri, three in Ggal and even seven in Apla.

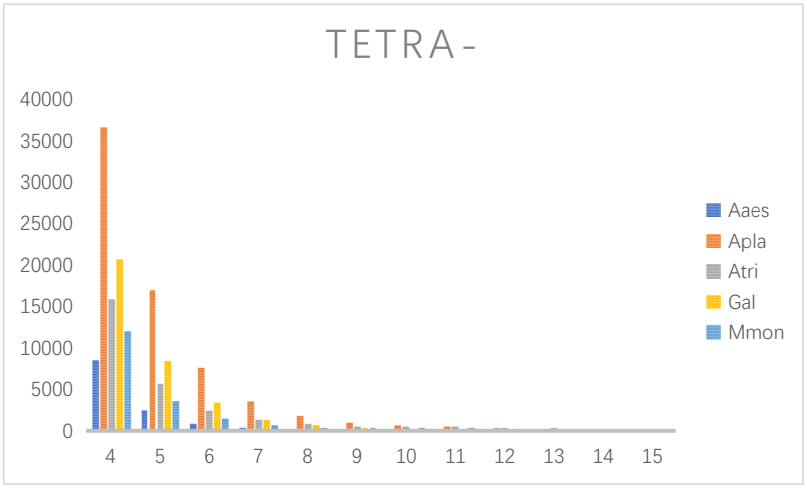

Fig. 5 Tetranucleotide repeat type microsatellite loci (repeats>15 partially obscure and omitted)

3.2.5 Pentanucleotide repeat type microsatellite loci

As shown in Figure 6, for Aaes and Apla, the most predominant pentanucleotide repeat type was (AAAAC)n , with frequencies of 0.52 loci/Mb and 1.94 loci/Mb, and accounted for 22.90% and 7.39% of the total pentanucleotide SSRs for Aaes and Apla, respectively. While for Mmon and Ggal the most frequent pentanucleotide repeat type was (GTTTT)n with frequencies of 0.35 loci/Mb and 0.77 loci/Mb, and accounted for 11.01% and 6.53% of the total pentanucleotide SSRs for Mmon and Ggal, respectively. However, for Atri (GGAAT) was the most dominant pentanucleotide repeat type with a frequency of 1.65 loci/Mb, accounting for 7.79% of the total. It can be seen that for pentanucleotide repeats, Atri, Apla and Ggal possess higher abundance than Aaes and Mmon. Besides, the maximum

number of pentanucleotide repeats among these five bird species was 450 (Ggal), while the second largest was 440 (Mmon), respectively.

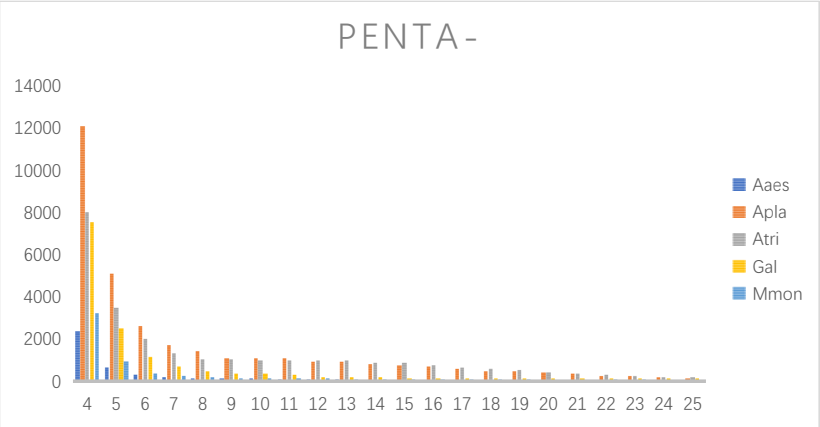

Fig. 6 Pentanucleotide repeat type microsatellite loci (repeats>25 partially obscure and omitted)

3.2.6 Hexanucleotide repeat type microsatellite loci

As shown in Figure 7, the repeat frequencies and densities of hexanucleotide SSRs were low compared with other SSR types. Most hexanucleotide repeats have a frequency of no more than 0.1 loci/Mb, but the frequency of (TAACCC)n in Atri is even as high as 4.35 loci/Mb with a density of 34.81 bp/Mb, and there are five repeats with a frequency of more than 2 loci/Mb, which is significantly higher than the frequency of other species. For the number of repeats, among the hexanucleotide SSRs Ggal exhibited an amazing 2770 repeats, while Mmon was also as high as 580 and Atri as high as 747.

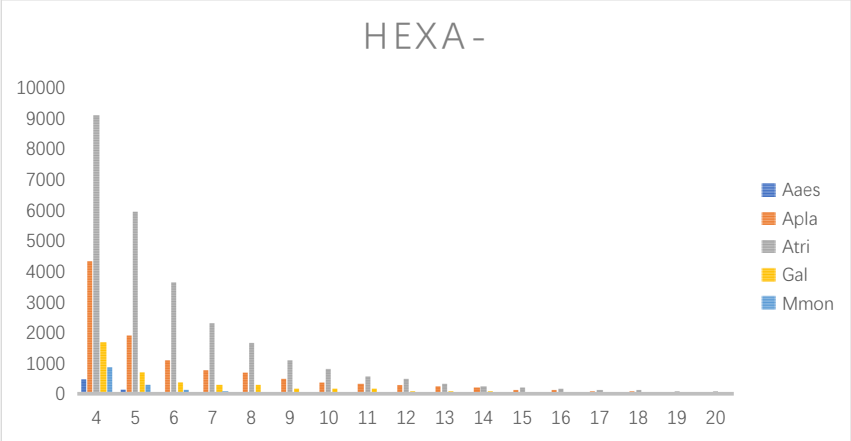

Fig. 7 Hexanucleotide repeat type microsatellite loci (repeats>20 partially obscure and omitted)

4. Summary

4.1. Similarity and variability of microsatellite distribution

In the present study, Amazona aestiva, Aaes; Myiopsitta monachus, Mmon; Acridotheres tristis, Atri; Anas platyrhynchos, Apla; Gallus gallus, Ggal were identified by MSDBv2.4.3 Microsatellites with 1-6 bp nucleotide motifs in the genome. We analyzed the number, relative frequency, relative density, number of repeats and GC content of microsatellites within the five species using the same parameter settings. Our results show that the two parrot species AA and MM are very different in SSR distribution despite being very similar in appearance, and also they are very different from other domesticated birds. We suggest that the distribution patterns of SSR in these birds may be unconservative and have little relationship with relatedness [4].

A total of 101,395, 171,161, 226,953, 458,231 and 272,487 P-SSRs were identified in the five selected bird species, with total frequencies of 2392.36 loci/Mb, 3726.92 loci/Mb, 209.93 loci/Mb, 385.54 loci/Mb and 258.69 loci/Mb, (fig. 1), representing 0.17%, 0.29%, 0.64%, 0.95% and 0.67% of the genome, respectively, and higher than cattle (0.48%) in all birds except two parrots (Baumann, Li et al. 2012), but except for ducks all macaques (0.83-0.88%) (Liu, Hou et al. 2017), and some of these differences may be caused by software differences during microsatellite searches, and differences in parameter settings (Sharma, Grover et al. 2007). It is reasonable to assume that the percentage of P-SSRs in the genome characterizes to some extent the complexity of the organism.

Among them, the most common types are all P-SSRs. for Aaes and Mmon ranked second are IP-SSRs and third are CD-SSRs, while the second most frequent in Apla and Ggal are CD-SSRs and the third most frequent are ICD-SSRs. while for Atri ranked second are ICD-SSRs and third are IP-SSRs. and among them The least among them all are CX-SSRs (Table 1). They correspond to each SSR with a large relative frequency difference, but their distribution in terms of type is somewhat related to affinity.

One of the characteristics of eukaryotic genomes is that there are more single nucleotide repeats than other nucleotide repeat categories (Sharma, Grover et al. 2007). In eukaryotes, there is a tendency for the number of SSRs to decrease as the length of repetitive unit sequences in SSRs increases (Sahu, Majee et al. 2020). However, it is clear that the percentage of various SSRs does not decrease with the increase in the length of individual SSR repeat unit sequences, but instead shows a trend of decreasing, then increasing and then decreasing in some species. We believe that this difference in microsatellite abundance order, which is only shown between vastly different species, reveals to some extent some very old evolutionary relationships and provides strong evidence that the emergence of SSRs is not random and that existing theories on SSR genesis are not sufficient to fully explain the generation and distribution of SSRs.

4.2. Number of repetitions of different types of microsatellites

The genomes of the five species all contained more A or T although the types of SSRs varied widely. in contrast, G- or C-rich repeats were rare in all types of SSRs. calculating the GC content of single to hexanucleotide perfect SSRs in the five genomes showed that the total GC content accounted for 1.95% (Aaes), 3.19% (Mmon) 29.71% (Atri), 16.73% (Apla) and 27.39% (Ggal). Indeed, this is also the case in other species, such as human (Subramanian, Mishra et al. 2003), Arabidopsis (Lawson and Zhang 2006) and brewer's yeast, and we suggest that this widespread phenomenon across species is related to the causes of microsatellite formation, namely point mutations (Messier, Li et al. . 1996) and replication slippage (Om Prakash and Seema 2010) are related to the fact that G-C structures are more stable and more difficult to form SSRs compared to A-T.

In the present study, we demonstrated that the distribution of six SSR repeats appeared significantly different in the genomes of five domesticated bird species, for single nucleotide SSRs, the Atri, Apla and Ggal genomes had significantly more SSRs with a high number of repeats than Aaes and Mmon, while microsatellites with 662 and 633 repeats even appeared in Ggal, for di- and trinucleotide SSRs, the distribution of the number of SSRs with high repeat counts in the genomes of the five species was similar, and when examining the repeat counts of the four-, five-, and six-nucleotide SSRs, we found that Ggal, Mmon, and Atri all had significantly more SSRs with high repeat counts in their genomes than the other species. And Atri was significantly higher in the type and frequency of high nucleotide repeats than other species, whether this suggests that Atri may have some important role in evolution that we have not discovered.

In summary, GC content and SSR repeat number may play an important role in most organisms. The results provide data to further investigate the role and interrelationship of GC content and SSR repeat number, revealing differences among the five bird species at the genomic level, and genetic diversity of species in domesticated birds. Our data will facilitate comparative genome mapping to understand the distribution and genomic structure of SSRs among these animal models and lay the foundation for further development and identification of more domesticated bird-specific SSRs.

References

[1] Huang J, Du LM, Li YZ, Li WC, Zhang XiuYue YUE (2012) Microsatellite distribution patterns in the whole genome of Hong Yuan chicken. Sichuan Journal of Zoology 31(3): 358-363.

[2] Abdelkrim J, Robertson BC, et al. (2009) Rapid and cost-effective development of species-specific microsatellite markers by genome sequencing. biotechniques 46(3): 185-192.

[3] Campregher C, Scharl T, et al. (2010) Nucleotide composition of microsatellites has implications for both replication fidelity and mismatch repair in human colorectal cells. hum Mol Genet 19(13): 2648-2657

[4] da Silva HE, Presti FT, et al. (2015) Development of microsatellite markers for hyacinth macaws (Anodorhynchus hyacinthinus) and their cross-amplification in other parrot species. bmc Res Notes 8: 736.

[5] Li LIU, Dybvig K, et al. (2000) The GAA trinucleotide repeat region regulates M9/pMGA gene expression in Mycoplasma gallisepticum. infect Immun 68(2): 871-876.