

Comparative Analysis of Microsatellite Sequences in Six Gamma-coronaviruses

Duo Zhang*, Haoxiang Wang

School of Life Science, Sichuan University, Chengdu, China

* Corresponding Author Email: Duo Zhang: zduo2023@163.com

Abstract. Microsatellite sequences are DNA or RNA repeat sequences with 1-6 bases as repeat units, also known as simple repeat sequences, that appear in the genome's non-coding, coding, and intergenic regions. Microsatellite sequences are extremely changeable and varied, and their presence encourages genomic variety and evolution. Gamma-coronavirus is a genus in the Coronaviridae subfamily. Microsatellites are extensively dispersed in gamma-coronavirus genomes' coding and non-coding regions. There are currently few papers on the microsatellite analysis of γ -coronaviruses. The distribution frequency, GC content, and nucleotide repeat sequences of γ -coronavirus SSR were studied in this work using MISA online program and SSRhunter, and three results were drawn: 1. Viruses have a short evolutionary period and a lower variety of microsatellites than prokaryotes and eukaryotes. 2. Type II microsatellites have a significant role in cross-species infection and gene expression differences. 3. Evolution might be skewed. This research will contribute to a better understanding of the structure, function, and evolutionary importance of microsatellites in tiny genomes.

Keywords: γ -coronavirus, Simple repeat sequences, Microsatellites, Comparative analysis of microsatellites

1. Introduction

Microsatellites, also known as simple sequence repeats, are short sequences composed of 1-6 bp long tandem repeats of motifs[1]. Microsatellite sequences have an extremely high mutation rate, which allows microsatellite to have extremely high genetic diversity[2]. Due to their mutability, microsatellite sequences can be characterized by a high mutation rate. Microsatellite is commonly used for forensic genetic identification[3]. Microsatellite is highly variable and varies greatly between individuals, possibly as seen found to be potentially associated with genome structure and function, so the emergence and fixation of microsatellites may contribute to genome diversity and evolution a result of strand slip during replication[4]. Mutations in microsatellites have been found to be potentially associated with genome structure and function, so the emergence and fixation of microsatellites may contribute to genome diversity and evolution[5]. In addition, microsatellite amplification has been confirmed to cause more than 30 neurological and neuromuscular diseases, such as epilepsy caused by abnormal expansions of TTTCa and TTTTA repeat sequences, Huntington's disease[6] caused by CAG repeat sequences, and is also associated with lung[7] cancer and gastric cancer[8,9]. Therefore, more and more researchers have paid great attention to the existence of microsatellites in different organisms.

Coronaviruses were first discovered from infected tissues of chickens in 1937. Due to the crown like spikes on the surface of the virus, it was named as "Coronavirus" by the International Committee on Nomenclature and Classification of Viruses. Coronaviruses belong to the family Coronaviridae and the order Nidovirales. Based on the viral morphology, genome structure and gene expression, Coronavirinae and Torovirinae were classified by the International Committee on the Classification of Viruses (ICTV) in 2012. Viruses of the subfamily Coronavirinae, commonly known as coronaviruses, They are divided into four gener.[9-10] The genus Ccov mainly includes coronaviruses that can infect birds, such as Infectious bronchitis virus (IBV), Turkey coronavirus, Duck coronavirus, Canada goose coronavirus, Beluga whale coronavirus and Bottlenose dolphin coronavirus (BdCoV), which can infect mammals.

Infectious bronchitis virus (IBV) is the causative agent of infectious bronchitis in chickens. In 1936, Beach and Schalm identified a virus as the causative agent of the disease, which was the first coronavirus identified and later named IBV[10]. Avian infectious bronchitis has been listed as a class B avian disease by the World Organization for Animal Health (OIE), and as a Class 2 animal epidemic disease in the catalogue of animal diseases issued by the Ministry of Agriculture of China.

The virus seriously affects the respiratory[11].

The Turkey coronavirus (TCoV) was first isolated in China in 2006 by Yang Geng et al. In the 1970s, it was identified as an important pathogen causing enteric diseases in turkey-turkey-induced diseases known as infectious gastroenteritis and coronavirus enteritis. TCoV can infect Turkey birds of any age, presenting with diarrhea, dehydration, growth arrest, weight loss, immune dysfunction, and abnormal egg production[12].

Duck coronavirus was discovered by Chen Guiqian et al. through high-throughput sequencing in 2012. It causes duck viral disease, an acute infectious disease characterized by severe diarrhea. Ducks around 20 days of age had the highest incidence rate, and the main clinical manifestation was acute onset Early loose stool, followed by diarrhea; White or yellow-green stools; The proboscis turn purple, and the epithelium is exfoliated and ulcerated[13].

The Canadian Goose coronavirus (CgCoV) was first proposed[14] goose coronavirus (GCoV) was first proposed[15]. However, the further antigenic characteristics of the virus have not been reported.

The complete genome sequence of beluga coronavirus SW1 (BWCoV SW1) was first published in 2008. The virus is host specific, and it is not possible to isolate the virus with non-homologous cell lines. Phylogenetic analysis of the viral genome showed that the virus was closely related to avian coronaviruses. Although the homology of its replicase gene was significantly lower than the critical point for species division, SW1 was still classified into the genus[16].

Bottlenose dolphin virus (BdCoV) is a novel coronavirus, designated as HKU22, detected from fecal samples of three apparently healthy Indo-Pacific bottlenose dolphins in Hong Kong, China between 2008 and 2010. Unlike belugas isolate SW1, HKU22 did not cause any obvious clinical symptoms, and bottlenose dolphins were only asymptomatic or mildly infected. In 2020, LeyiWang et al. found that a group of dolphins infected with BdCoV had the main clinical reactions of lethargy, loss of appetite and diarrhea[17].

After more and more biological genes have been sequenced, researchers have found that there are always repetitive sequences in genes. For example, microsatellite sequences account for about 3% of the human genome[18], much more than protein coding sequences. Previous studies have shown that the occurrence of repeat sequences is not random. Since most microsatellite sequences are distributed in the non-coding regions of the genome, it is difficult to analyze their connotation by traditional experimental methods.

IBV is a representative species of γ -coronavirus and the first coronavirus discovered in human history. Infectious bronchitis caused by IBV has been listed as a Class II animal disease in China.

Gamma-coronaviruses have diverse hosts, and differences in their evolution are not fully understood. Previous studies[19-20] on HIV-1 and HCV have suggested that microsatellites may have potential contributions to viral evolution. In plant viruses, microsatellites may contribute to secondary structure formation[21]. Studies on bacterial microsatellites have shown that microsatellites can contribute to the rapid evolution of new species[22]. Based on previous studies, we hypothesized whether γ -coronaviruses have a similar mechanism of microsatellites-related evolution. In this study, we extracted the number, type, relative abundance and relative density of microsatellites from six gamma-coronaviruses. By analyzing the correlation between various microsatellites, we further explored the potential genetic characteristics of gamma-coronaviruses and provided new ideas for researchers to prevent and treat viral animal diseases.

2. Materials and Methods:

2.1. Genome sequence data

The complete genome sequences of 6 γ -coronaviruses with complete sequence annotations were queried and downloaded from the Genbank database of NCBI, and the general genome information is shown in Table 1.

Table 1. Genomic information of the six gammoviruses

Name	Accession Number	Genome Size(nt)	Host
Infectious bronchitis virus	GCA_000862965.1	27608	Chicken
Turkey coronavirus	GCA_000880055.1	27,657	Turkey
Duck coronavirus	GCA_012271565.1	27754	Duck
Canada goose coronavirus	GCA_012271745.1	28539	Canada goose
Begula whale coronavirus SW1	GCA_000872845.1	31,686	Begula whale
Bottlenose dolphin coronavirus HKU22	GCA_029883535.1	31769	Bottlenose dolphin

2.2. Extraction of microsatellites

Based on the six gamma-virus genome sequences obtained from Genbank, the MISA online software[23] was used to search type I microsatellite and set the repeat number ≥ 6 . SSRHunter software was used to search microsatellite loci with nucleotide repeat numbers of 2,3,4,5 and 6, and the corresponding repeat numbers were set as 3,3,3,3 and 3, respectively. For example, if a single base A is repeated 6 times in a row, it can be considered to form a type I microsatellite sequence, which can be denoted as (A)₆. The two AT bases were repeated three times to form a type II microsatellite sequence, which can be denoted as (AT)₃. Type III microsatellites and above are defined in the same way, and other software parameters are set to default values. The results of the two software were combined to calculate the types and quantities of types 1 to 6 microsatellites in the six kinds of γ -coronaviruses.

2.3. Statistical analysis of microsatellites

After microsatellite extraction, Excel was used to preprocess the data, and the cyclic sequences were included in the same type of SSR. For example, the cyclic sequence of ACG contains ACG, AGC, GAC, GCA, CGA and CAG, and the above six sequences are considered to be the same microsatellite sequence after data preprocessing.

3. Results

3.1. GC content statistics

All the six viruses belong to the genus γ coronavirus, which can infect different hosts, and their whole genomes are highly similar. Statistical analysis showed that the GC content of the genomes of the six γ -coronaviruses was almost inhibited, and the GC content of the six genome sequences was between 38.25% and 39.27%. Among them, IBV and Duck coronavirus had the largest GC content, and Turkey coronavirus had the smallest GC content.

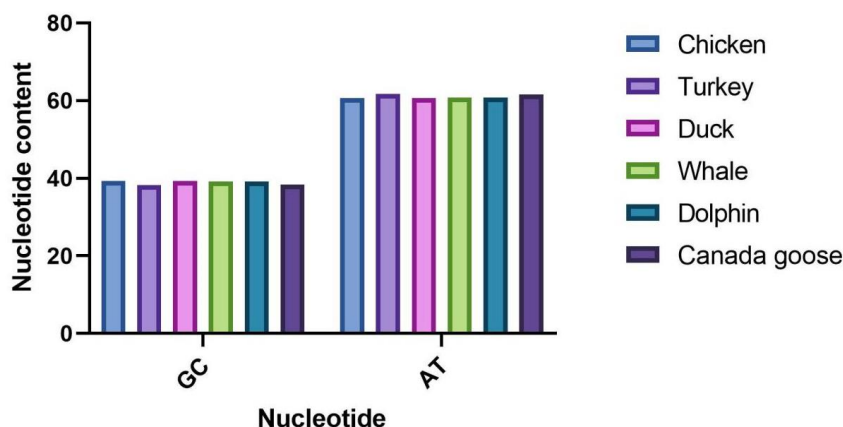


Fig. 1 Statistics of the GC content of the genomes of the six γ -coronaviruses

According to the proportion of GC in the genome, it is difficult to obtain the difference information of the six kinds of γ -coronaviruses from the perspective of base composition. However, microsatellites exist in multiple places in the genome and gene mutations, and the change may only be the decrease or increase in the number of microsatellite repeat units. It maybe difficult to find the cross-species transmission ability of γ -coronavirus caused by microsatellite changes from the perspective of base composition proportion, which is why this study focused on microsatellites.

3.2. The number of microsatellite statistics

The total number of microsatellites in the genomics of the six γ coronaviruses ranged from 96 to 110, with Turkey coronavirus being the least abundant and Canada goose coronavirus being the most abundant. Type 2 microsatellite was the most common type, followed by type 3 and type 1 microsatellite. No microsatellites of type 4 or above were found.

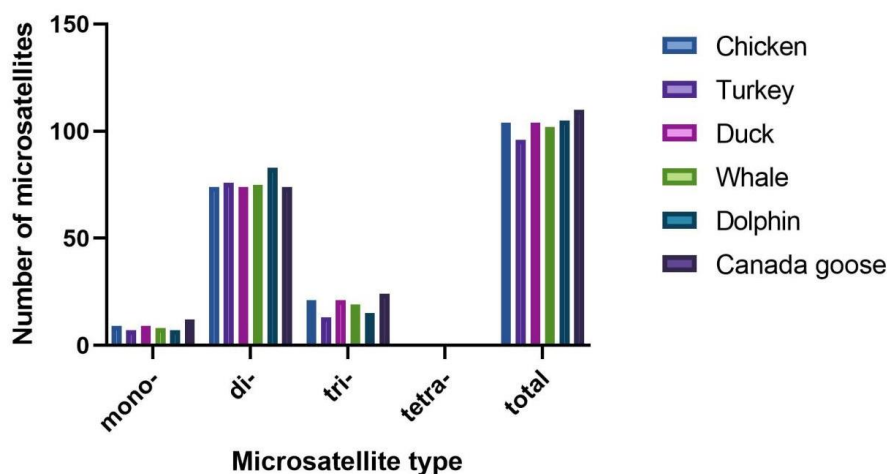


Fig. 2 The number of microsatellites in the genomes of the six gamma-coronaviruses

The difference in the number of different types of microsatellites highlights the important role of type II microsatellites in the evolution of γ -coronaviruses. The hosts of γ -coronaviruses varied from chickens, Turkey, ducks, and Canada goose (Orniths) to whales and dolphins (Mammalia). The average number of type II microsatellites increased, which may be due to the formation of a large number of microsatellites in the natural host of γ -coronaviruses, chicken, through long-term evolution and mutation, and the rapid adaptation to new hosts by increasing the number of repeats during cross-host evolution.

After that, the relative abundance and density of microsatellites were studied. Relative abundance = (number of microsatellites/genome length) * 1kbp, expressed in 1/kb, represents the average number of microsatellites per 1000bp in the genome. Relative density = (microsatellite sequence length/genome length) * 1kbp, in bp/kbp, represents the average number of bases of microsatellite sequences present per 1000bp in the genome. Both relative abundance and relative density can be

used to normalize the number of microsatellite and sequence length, and eliminate the influence caused by the difference of genome length.

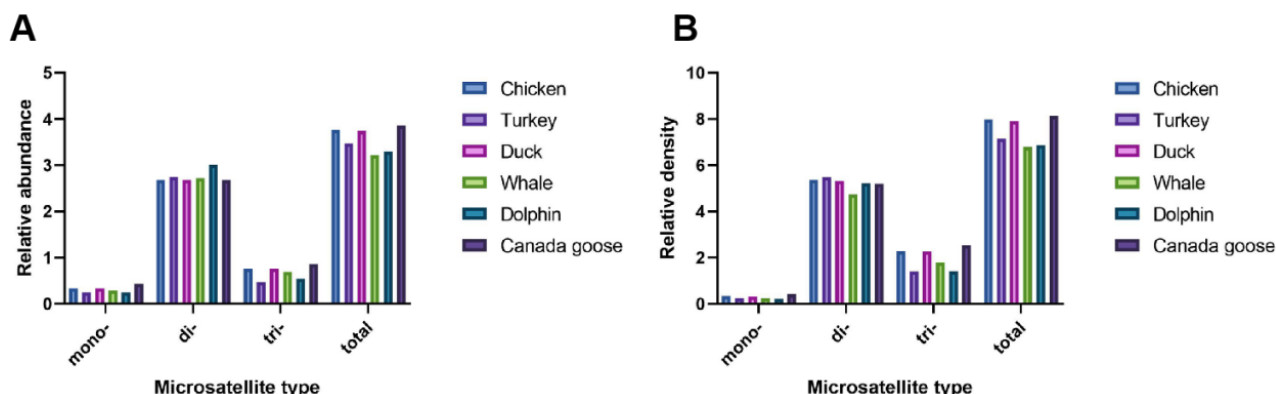


Fig. 3 Relative abundance, relative density of microsatellites in the genomes of six gammoviruses

3.3. Statistical analysis of different types of microsatellites

3.3.1 Statistics of type 1 microsatellites

Statistical analysis of type 1 microsatellite in the genomes of six gamma-coronaviruses showed similar repeat type distribution results. Canada goose coronavirus had the highest number of type I microsatellites (11), while Turkey coronavirus and Dolphin coronavirus HKU22 each had the lowest number of type I microsatellites (7). Among them, T repeats accounted for the largest proportion, occurring 25 times in the six viral genomes, followed by A repeats, occurring 23 times, C and G repeats were the least, occurring 2 and 1 times, respectively. G repeats were only found in Canada goose coronavirus, and there were a large number of T repeats. C repeats were only found in IBV and Duck coronavirus. Beluga whale coronavirus had more A repeats and less T repeats. From the perspective of Ornithoptera and Mammalia, the diversity and average number of type 1 microsatellites of avian gammoviruses were higher than those of mammalia gammoviruses.

The difference of microsatellite type 1 among the six groups of viruses suggests that it may be closely related to the survival of viruses. If there is a large change in microsatellite type I, the virus may not be able to perform normal physiological functions.

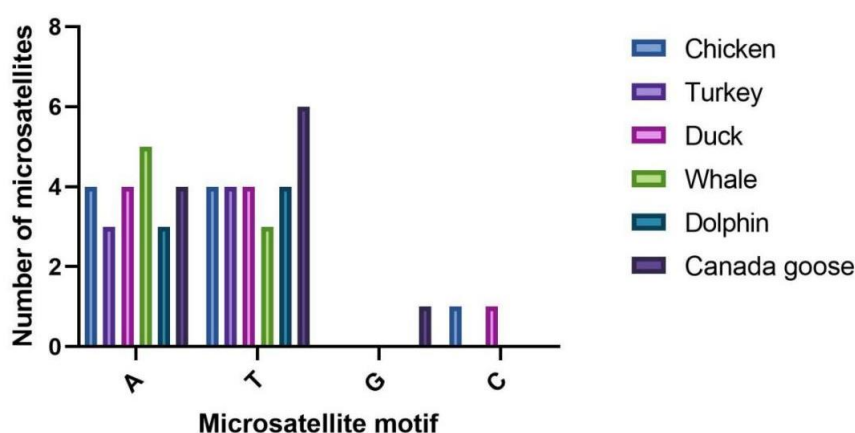


Fig. 4 Type 1 microsatellite distribution of six different types of gamma-coronaviruses

3.3.2 Number statistics of type II microsatellites

Dolphin coronavirus contained 83 type 2 microsatellites, the highest number of type 2 microsatellites, and the other five types were in the range of 74 to 76, with a small difference. All the six viruses had the highest number of TG/GT(183) and AT/TA(100) repeats, followed by AG/GA(67), AC/CA(63), CT/TC(35) and CG/GC(9) repeats. There were significant differences among different species. Canada goose coronavirus did not have CG/GC microsatellite but had more CT/TC microsatellite. From the perspective of birds and mammals, AC/CA microsatellite types of γ -

coronaviruses in birds were significantly more than those in mammals, and TG/GT microsatellite types of γ -coronaviruses in mammals were significantly more than those in birds.

The difference of type 2 microsatellites among six groups of viruses indicated that type 2 microsatellites might be closely related to the transmission and evolution of viruses. The changes of physicochemical properties and cellular environment in the host may lead to the changes of the number and type of microsatellites, through which the virus ADAPTS to the new environment.

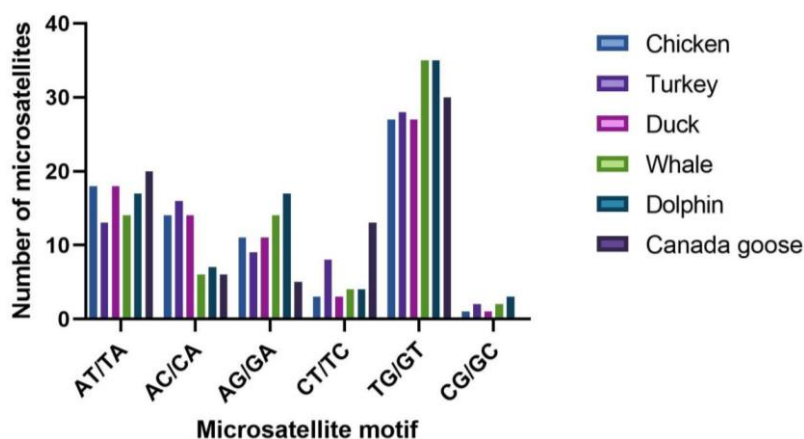


Fig. 5 Distribution of microsatellites of six groups of γ -coronaviruses type 2

3.3.3 Statistics of three types of microsatellites

TGG/TTG and ATG type 3 microsatellites were the most frequent among the three types of microsatellites. With the increase of the length and complexity of the microsatellite sequence, the distribution of three types of microsatellite and the number of the same type of γ -coronaviruses in the six species are not the same. Especially, there are obvious differences between avian and mammalian γ -coronaviruses in some three types of microsatellite, such as avian γ -coronaviruses unique CTT/CCT microsatellite. TGG/TTG microsatellite was significantly more than that of mammalian γ CoVs, while ACG and AGG/AAG microsatellite of mammalian γ CoVs were significantly more than those of ornithoviruses.

The number and types of three types of microsatellites vary greatly among different species, which maybe because the three types of microsatellites provide a rich gene pool for γ -coronaviruses, thereby enabling them to acquire the ability of rapid cross-host transmission.

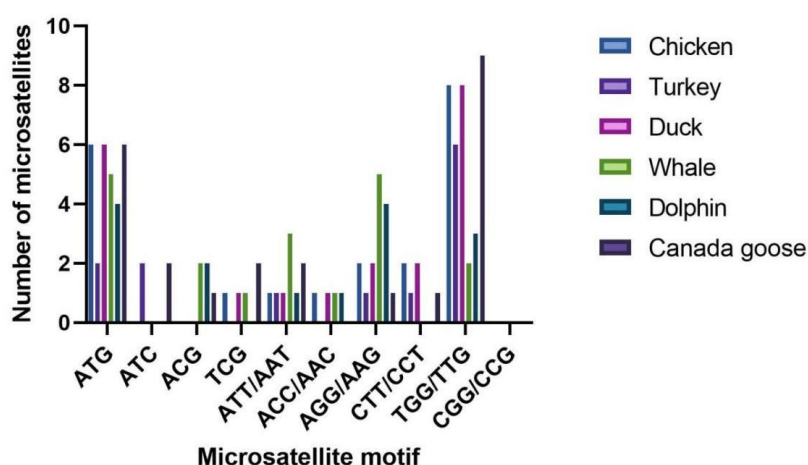


Fig. 6 Distribution map of three types of microsatellites of six groups of gamma-coronaviruses

4. Discussions

In this study, we identified Infectious bronchitis virus through MISA online software with SSRhunter; Turkey coronavirus; Duck coronavirus; Begula whale coronavirus; Bottlenose dolphin

coronavirus. The Canada goose coronavirus genome aggregates microsatellites with 1-6bp nucleotide motifs, namely type I-VI microsatellites. But only type 1-3 microsatellite was detected, which was similar to other reports on viral microsatellite. At present, the longest microsatellite sequence found in viruses is microsatellite type 5, and the number of microsatellite type 4 and microsatellite type 5 is small[24]. Based on the analysis results, we can reach the following conclusions: 1. The degree of sequence similarity is high, and the length of microsatellite in SARS-CoV-2 is generally short. Different from eukaryotes, the relative abundance, density, sequence length and microsatellite types of viral microsatellite are lower than those of eukaryotic microsatellite, which indicates that the evolution time of virus is short, and the diversity of microsatellite is not as good as that of prokaryotic and eukaryotic microsatellite. 2. Although the degree of similarity is high, the distribution of microsatellite among species is still different, and the difference is more obvious within the same type. The type and number of microsatellite in different species were significantly different, indicating that microsatellite type 2 may play a key role in cross-species infection and differential gene expression. 3. There were few C/ G-containing microsatellites, a few C/ G-containing microsatellites in type I and type II microsatellites, and no C/ G-containing microsatellites in type III microsatellites. Recent studies have shown that this phenomenon does not occur in eukaryotes, prokaryotes, and animal and plant viruses[25-26], indicating that evolution may be biased.

The same parameters were used to analyze the GC content, type, number, species, relative abundance and relative density of microsatellites for the six gammaviruses. The results showed that although IBV and TCoV infected in different hosts, the expression of microsatellites in their genomes was identical. Although all the six viruses are γ -coronaviruses, the distribution of microsatellite among viruses is quite different. Studies suggest that the evolution of viruses during cross-host transmission leads to the non-conservation of microsatellite sequences due to different hosts. Among them, the difference of type 1 microsatellite is relatively small. We believe that type 1 microsatellite may play an important physiological role in γ -coronaviruses, and a large number of type 1 microsatellite mutations may lead to the defect of individual physiological function of the virus so that it cannot survive. The difference of type 2 microsatellites was significant, suggesting that type 2 microsatellites might play an important role in the cross-host transmission of γ viruses. Type 3 microsatellites also have large differences among species, so we speculate that type 3 microsatellites can play a role in enriching the gene bank reserve, and the scope of research can be further expanded to conduct a wide range of studies on the whole coronavirus.

In the microsatellite studies of six kinds of gamma-coronaviruses, the viruses can be divided into avian-class gamma-coronaviruses and mammalia gamma-coronaviruses according to their hosts. There was no significant difference in type I microsatellite between the two groups, but in type II and type III microsatellite, the viruses infecting the two different hosts showed significant differences in some specific gene sequences. There were more AC/CA microsatellite types in ornithoviruses than in mammalia, and more TG/GT microsatellite types in mammalia than in ornithoviruses. CTT/CCT microsatellite types were unique to ornithoviruses, and TGG/TTG microsatellite types were significantly more than that in mammalia. However, the number of ACG and AGG/AAG microsatellites in mammalian γ -coronaviruses was significantly higher than that in avian viruses. This has also been reported in previous studies, for example, SARS-CoV strains that infect different hosts also have different microsatellite species. In general, this study also provides ideas for further studies on the number and evolution of microsatellite species.

References

- [1] Schlötterer C. Evolutionary dynamics of microsatellite DNA[J]. *Chromosoma* 109, 365–371 (2000).
- [2] Bhargava A, Fuentes F.F. Mutational Dynamics of Microsatellites[J]. *Mol Biotechnol* 44, 250–266 (2010).
- [3] Haddrill, Penelope R. Developments in forensic DNA analysis[J]. *Emerging topics in life sciences*, 2021, 5(3).

- [4] Schmidt, Monika H M, Pearson, et al. Disease-associated repeat instability and mismatch repair[J]. DNA repair, 2015, 38:117-26.
- [5] Haasl, Ryan J, Payseur, et al. Microsatellites as targets of natural selection[J]. Molecular biology and evolution, 2012, 30(2):285-98.
- [6] Hiroyuki Ishiura, Koichiro Doi, Jun Mitsui, et al. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy[J]. Nature genetics, 2018, 50(4):581-590.
- [7] Nakamori Masayuki, Panigrahi Gagan B, Lanni Stella, et al. Terence. A slipped-CAG DNA-binding small molecule induces trinucleotide-repeat contractions in vivo[J]. Nature genetics, 2020, 52(2):146-159.
- [8] Wadowska Katarzyna, Bil-Lula Iwona, Trembecki Łukasz, et al. Genetic Markers in Lung Cancer Diagnosis: A Review[J]. International journal of molecular sciences, 2020, 21(13).
- [9] Ratti Margherita, Ratti Margherita, Lampis Andrea, et al. Microsatellite instability in gastric cancer: molecular bases, clinical perspectives, and new treatment approaches. Cellular and molecular life sciences : CMLS, 2018, 75(22):4151-4162.
- [10] Owens, Flores, Di Serio, et al. (2012). Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses[C].
- [11] Liu F. Isolation, identification, biological characteristics and preliminary study on vaccine candidate strains of infectious bronchitis viruses [D]. South China Agricultural University,2019. (in Chinese)
- [12] Yang Xinxin, Zhao Lihong, Qiao Jian, et al. Isolation and preliminary identification of Turkey coronavirus [J]. Chinese Journal of Animal Science and Veterinary Medicine,2006(11):1241- 1244.
- [13] Chen Gui-Qian, Zhuang Qing-Ye, Wang Kai-Cheng, et al. Identification and survey of a novel avian coronavirus in ducks[J]. PloS one, 2013, 8(8).
- [14] Papineau Amber, Papineau Amber, Berhane Yohannes, Wylie, Todd N., et al. Genome Organization of Canada Goose Coronavirus, A Novel Species Identified in a Mass Die-off of Canada Geese[J]. Scientific reports, 2019, 9(1):5954.
- [15] Jonassen CM, Kofstad Tone, Larsen Inger-Lise,et al. Molecular identification and characterization of novel coronaviruses infecting graylag geese (*Anser anser*), feral pigeons (*Columbia livia*) and mallards (*Anas platyrhynchos*)[J]. The Journal of general virology, 2005, 86(Pt 6):1597-1607.
- [16] Mihindukulasuriya Kathie A, Wu Guang, St Leger Judy,et al. Identification of a novel coronavirus from a beluga whale by using a panviral microarray[J]. Journal of virology, 2008, 82(10).
- [17] Wang Leyi, Maddox Carol, Terio Karen, et al. Detection and Characterization of New Coronavirus in Bottlenose Dolphin, United States, 2019[J]. Emerging infectious diseases, 2020, 26(7):1610-1612.
- [18] Li You-Chun, Korol Abraham B, Fahima Tzion,et al. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review[J]. Molecular ecology, 2002, 11(12):2453-65.
- [19] Zhao Xiangyan, Tian Yonglei, Yang Ronghua,et al. Coevolution between simple sequence repeats (SSRs) and virus genome size[J]. BMC genomics, 2012, 13.
- [20] Chen Ming, Tan Zhongyang, Zeng Guangming. Microsatellite is an important component of complete hepatitis C virus genomes[J]. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 2011, 11(7):1646-54.
- [21] Qin Lü, Zhang Zhixiang, Zhao Xiangyan,et al. Survey and analysis of simple sequence repeats (SSRs) present in the genomes of plant viroids[J]. FEBS open bio, 2014, 4.
- [22] Moxon E R, Rainey P B, Nowak M A,et al. Adaptive evolution of highly mutable loci in pathogenic bacteria[J]. Current biology : CB, 1994, 4(1):24-33.
- [23] Beier S, Thiel T, Münch T, et al. Mascher M (2017) MISA-web: a web server for microsatellite prediction[J]. Bioinformatics 33 2583–2585.
- [24] Jiang Shuai, You Changqiao, Ding Yunyun, et al. Study on the genetic diversity of 2019-ncov based on microsatellite association test [J]. Journal of Laser Biology,2023,32(03):208-216.
- [25] Alam CM, Singh Avadhesh Kumar, Sharfuddin Choudhary, et al. Incidence, complexity and diversity of simple sequence repeats across potexvirus genomes[J]. Gene, 2014, 537(2).
- [26] Alam CM, Singh Avadhesh Kumar, Sharfuddin Choudhary, et al. Genome-wide scan for analysis of simple and imperfect microsatellites in diverse carlaviruses[J]. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 2013, 21.