Distribution Characteristics of DNA Complementary Palindrome Patterns in the African Swine Fever Virus Genome

Authors: high mountain

Date: 2018-08-29T00:00:00+00:00

Abstract

Inspired by the discovery of complementary palindromic small RNAs, this study considers the biological function of DNA complementary palindromic patterns at the small RNA level and reveals the distribution characteristics of longer (14 bp and above) DNA complementary palindromic patterns in the African swine fever virus genome. The DNA complementary palindromic patterns in the African swine fever virus genome disclosed in this study can be used to design primers or probes to improve the sensitivity and specificity of virus detection; they can also be directly used to design small interfering RNA (siRNA) for RNA interference experiments. Thus, without the need for viral infection, more basic researchers can use this indirect method to study the infection and pathogenic mechanisms of African swine fever virus. The ideas and methods provided by this study for investigating DNA complementary palindromic patterns in viral genomes can be extended to other microorganisms or to the study of gene function or evolution in plants and animals, and have important theoretical significance.

Full Text

Title

Study on the Distribution Characteristics of DNA Complementary Palindromic Patterns in the African Swine Fever Virus Genome

Authors: Xu Xiaofeng , Chen Ze , Luo Jianxun , Liu Guangyuan , Ren Qiaoyun , Luo Jin , Yin Hong , *Gao Shan* **Affiliations:** State Key Laboratory of Veterinary Etiological Biology, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences; Gansu Provincial Key Laboratory of Animal Parasitic Diseases; Jiangsu Co-innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Lanzhou

730046; College of Life Sciences, Nankai University, Tianjin 300071 **Correspondence:** gao_{shan}@mail.nankai.edu.cn; yinhong@caas.cn **Funding:** Basic Research Fund of Central Public-interest Scientific Institution

Abstract

Inspired by the discovery of complementary palindromic small RNAs, this study examined the biological functions of DNA complementary palindromic patterns at the small RNA level and revealed the distribution characteristics of long (14 bp and above) DNA complementary palindromic patterns in the African Swine Fever Virus (ASFV) genome. The DNA complementary palindromic patterns identified in the publicly available ASFV genome can be used to design primers or probes to improve the sensitivity and specificity of viral detection, or directly employed to design small interfering RNAs (siRNAs) for RNA interference experiments. This approach enables more basic researchers to investigate ASFV infection and pathogenic mechanisms without requiring actual viral infection. The methodology and analytical framework for studying DNA complementary palindromic patterns in viral genomes can be extended to other microorganisms or to plant and animal gene function and evolution studies, holding significant theoretical importance.

Introduction

Palindromic sequences, also known as palindromic DNA motifs, are inverted repeat sequences widely present in various organism genomes. Known biological functions of DNA palindromic patterns include restriction enzyme sites, methylation sites, and T cell receptor-related sequences [1]. Unlike the classical definition that requires identical sequences on both DNA strands from 5' to 3' direction, this study redefines DNA palindromic patterns and DNA complementary palindromic patterns. Inspired by restriction enzyme site patterns, current research on DNA complementary palindromic patterns has focused primarily on shorter motifs (generally no more than 10 bp) under strict definitions where all bases participate in complementary pairing, leaving the biological functions of large DNA palindromic and complementary palindromic patterns abundant in plant and animal genomes largely unknown. In 2018, Gao Shan from Nankai University and Chen Ze from Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, first reported the existence of complementary palindromic small RNAs (cpsRNAs) in SARS coronavirus (SARS-CoV) and, through combined evolutionary and molecular functional analysis, preliminarily demonstrated that cpsRNAs may play roles in SARS-CoV infection and pathogenesis [1]. This study was the first to consider the biological functions of DNA complementary palindromic patterns at the small RNA level and revealed important characteristics of long (14 bp and above) DNA complementary palindromic patterns in viral genomes (see Results). These characteristic analyses provide new ideas and methods for studying viral infection and pathogenic mechanisms. African Swine Fever Virus (ASFV [2]) was first identified in China in

2018 and causes an acute, highly contagious swine fever with mortality rates up to 100% [3], necessitating urgent research on its infection and pathogenic mechanisms. Applying the concepts and methods from our previous work [1], this study conducted statistical analysis of DNA complementary palindromic pattern characteristics in the ASFV genome. We found that the genomic density of DNA complementary palindromic patterns in ASFV is significantly higher than in other double-stranded DNA viruses, and some patterns are highly similar to those functionally validated in the SARS-CoV genome. These findings provide novel approaches for future ASFV infection and pathogenic mechanism research.

1. Methods and Data

The reference genome sequences of eight viruses (HPV-18, HBV, HCV, HIV-1, EBV, SMRV, SARS-CoV, and ASFV; accession numbers KU298886.1, JQ688404.1, D11168.1, KM390026.1, M80517.1, M23385.1, DQ497008.1, and FN557520.1) were obtained from the NCBI GenBank database. The human GAPDH gene sequence (ENSG00000111640.14) was retrieved from the Ensembl genome database. ASFV small RNA high-throughput sequencing data (SRA: ERP018944) were downloaded from the NCBI SRA database. Data quality control for high-throughput sequencing was performed using Fastq_{clean} v2.0 [4]; sequence alignment to viral reference genomes used Bowtie v0.12.7; statistical analysis and plotting employed R v2.15.3 [5]; and alignment result verification utilized Tablet v1.15.09.01 [6]. Minimum Free Energy (MFE) calculations for DNA complementary palindromic pattern secondary structures were performed using the online RNAfold service (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). The intramolecular annealing temperature (Tm) of stem-loop structures was calculated using the formula 4(G+C)+2(A+T). When counting DNA palindromic or complementary palindromic patterns in viral genomes, we required a stem length of at least 7 bp and a loop length of 0-5 bp. Genomic density of DNA complementary palindromic patterns was calculated as: (total number of patterns in a genome $\times$ 1000) / genome length.

We redefined DNA palindromic patterns and DNA complementary palindromic patterns based on two considerations: first, preliminary studies suggested they have different biological significance; second, they exhibit distinct distribution characteristics in animal virus genomes. DNA palindromic patterns require complementary pairing between the two DNA strands in the 5' to 3' direction (e.g., ATCGGCTA), while DNA complementary palindromic patterns follow the classical palindromic definition (e.g., ATCGCGAT). In 2017, Bu Wenjun and Gao Shan from Nankai University first reported the full-length sequences of two long non-coding RNAs (lncRNAs) transcribed from the human mitochondrial D-loop region and the first discovered palindromic small RNA (psRNA), speculating that DNA palindromic patterns of small RNA length might be involved in transcriptional regulation [7]. In 2015, Gao Shan and colleagues serendip-

itously discovered a series of cpsRNAs derived from SARS-CoV during small RNA high-throughput sequencing studies of viral signature sequences. These sequences originate from the DNA complementary palindromic pattern TCTTTAACAAGCTTGTTAAAGA (see Figure 1 [Figure 1: see original paper]A). Subsequent RNA interference experiments with cpsRNA SARS-CoV-cpsR-19 (see Figure 1A) demonstrated that this small RNA could induce significant apoptosis. Due to their unique sequence patterns, psRNAs and cpsRNAs are defined as novel small RNA classes, though they lack biological association. Since DNA complementary palindromic patterns may be related to viral infection or pathogenesis, this study focuses exclusively on DNA complementary palindromic patterns in viral genomes, with no further discussion of DNA palindromic patterns.

Figure 1. Complemented palindromic DNA motif and complemented palindromic small RNA. A. Complementary palindromic small RNAs discovered in SARS-CoV (GenBank: DQ497008.1) with lengths of 18, 19 (named SARS-CoV-cpsR-19), and 21 nt. B. Redefined DNA complementary palindromic pattern (sequence from Table 1 , entry 2). C. DNA complementary palindromic patterns from SARS-CoV (GenBank: DQ497008.1, left) and ASFV (GenBank: FN557520.1, right), from Table 1 entries 1 and 2. The RNA secondary structure calculated from a DNA complementary palindromic pattern is termed the secondary structure of that DNA pattern. This structure does not represent the actual secondary structure of the DNA sequence itself; calculated properties such as minimum free energy belong to the corresponding RNA secondary structure, thus T in sequences can be represented as U.

Current research on DNA complementary palindromic patterns has focused primarily on shorter motifs (generally no more than 10 bp) under strict definitions where all bases participate in complementary pairing, which presents significant limitations. Chew et al. first reported the characteristics of DNA complementary palindromic patterns in the SARS-CoV genome [8], with two main findings: (1) 4 bp DNA complementary palindromic patterns showed significantly lower frequency across all analyzed coronavirus genomes; (2) 6 bp patterns were significantly less frequent only in SARS-CoV (not all coronaviruses). The study concluded that the scarcity of 6 bp DNA complementary palindromic patterns in SARS-CoV might help the virus evade certain host defense mechanisms. Two longer patterns (14+ bp) were also identified: TCTTTAACAAGCTTGTTAAAGA and TAAAATTAATTTTA. Probability modeling suggested these were not random occurrences. However, using the classical definition, Chew et al. found only two long DNA complementary palindromic patterns in SARS-CoV. Inspired by cpsRNA discoveries, we relaxed the definition requirements and identified 27 additional long DNA complementary palindromic patterns in SARS-CoV [7].

Referencing RNA hairpin definitions, we divided DNA palindromic and complementary palindromic patterns into two sequence components: the stem that forms the palindrome or complement, and the loop that does not participate

(see Figure 1B). We defined the RNA secondary structure calculated from a DNA complementary palindromic pattern as the pattern' s secondary structure. This structure does not represent the actual DNA secondary structure; calculated properties like minimum free energy belong to the corresponding RNA secondary structure. Stem-loop definitions in DNA palindromic and complementary palindromic patterns are DNA sequence-based and independent of secondary structure; not all stem bases necessarily participate in base pairing in the calculated secondary structure. MFE calculations for DNA complementary palindromic patterns follow the same principles as RNA hairpin MFE calculations, considering non-canonical GU pairing (displayed as GU or GT). Tm calculations for stem-loop structures use the formula 4(G+C)+2(A+T) applied to all bases in one stem side (see Table 1), independent of the corresponding secondary structure.

After relaxing requirements, DNA complementary palindromic patterns could contain loops of up to 5 bases (this parameter applies only to viral genomes). Statistical analysis of eight mammalian virus genomes revealed three typical features of long (14+ bp) DNA complementary palindromic patterns: (1) pattern quantity decreases significantly with length, with patterns beyond a certain length (31 bp) absent, demonstrating a truncation effect; (2) only an extremely small number of patterns have corresponding cpsRNAs detectable in current databases, indicating low probability of cpsRNA generation (possibly due to technical detection limitations); (3) some patterns are highly evolutionarily conserved with only a few transition mutations that do not affect secondary structure stability. To compare DNA complementary palindromic pattern content across viral genomes, we defined genomic density (see Methods and Data). Significant differences were observed both between different viruses and among subtypes/strains of the same virus. The genomic densities for HPV-18 (dsDNA), HBV (dsDNA), HCV (+ssRNA), HIV-1 (+ssRNA), EBV (dsDNA), SMRV (ssRNA), SARS-CoV (+ssRNA), and ASFV (dsDNA) were 0.64 $(1000\$\times5/7857), 0.93(1000\times3/3215), 0.95(1000\times9/9436), 0.21(1000\times2/9709), 0.62(1000\times115/184113), 0.68(100$ $ssRNAviruses, HIV-1showedsignificantlylowergenomicdensitythanothervirusesandevenlowerthanhuma$ Among dsDNA viruses, densities varied substantially from HPV-18 to HBV to ASFV. ASFV exhibited the highest genomic density at 1.1, with 199 DNA complementary palindromic patterns, 16 of which (see Table 1) were highly similar to those found in SARS-CoV (see Figure 1C).

Table 1. Sixteen DNA complementary palindromic patterns in the ASFV genome. The table lists patterns with their start positions, lengths, and sequences. The asterisk indicates a pattern from SARS-CoV (GenBank: DQ497008.1) that encodes the first functionally validated cpsRNA, SARS-CoV-cpsR-19, which induces significant apoptosis [1]. The 16 ASFV patterns (GenBank: FN557520.1) share similar properties with SARS-CoV-cpsR-19. Tm represents the intramolecular annealing temperature of the stem-loop structure; MFE represents the minimum free energy of the secondary structure.

For statistical analysis of viral genome DNA complementary palindromic pat-

terns, we examined four attribute distributions: length, GC content, Tm, and MFE. Length distributions in three viruses (EBV, SARS-CoV, and ASFV) generally showed significant decreasing trends with length, with patterns beyond 31 bp absent (see Figure 2 [Figure 2: see original paper]A). SARS-CoV peaked at 17 bp patterns; EBV showed anomalous increases at 19 bp and 23 bp, leading to the discovery of an EBV-encoded microRNA [9]. GC content distributions represented three types: high (EBV), medium (SARS-CoV), and low (ASFV) (see Figure 2B). Tm distributions generally concentrated at 18°C and 20°C across the three viruses (see Figure 2C), with ASFV anomalies at 14°C and 16°C and EBV anomalies at 24°C and 30°C. The 20°C temperature matches small RNA library preparation conditions, where adapters may fail to ligate due to cpsRNA secondary structures, causing substantial cpsRNA loss in small RNA high-throughput sequencing. MFE distributions also represented three types: high (ASFV), medium (SARS-CoV), and low (EBV). These distributional characteristics reflect important intrinsic viral properties significant for viral sequence analysis.

Figure 2. Distributive characteristics of complemented palindromic DNA motifs. A. Length distribution of DNA complementary palindromic patterns in three viral genomes. B. GC content distribution. C. Intramolecular annealing temperature distribution.

3. Discussion and Conclusion

This study examined the biological functions of DNA complementary palindromic patterns at the small RNA level and revealed important characteristics of long (14+ bp) patterns in the ASFV genome. Although no cpsRNAs from ASFV DNA complementary palindromic pattern regions were detected in the single available small RNA high-throughput sequencing dataset from NCBI SRA (ERP018944), our analytical results still support the relevance of these patterns to ASFV infection and pathogenic mechanisms. The significance of this study includes: providing a methodology for investigating DNA complementary palindromic patterns in viral genomes; enabling primer or probe design based on pattern conservation to improve detection sensitivity and specificity; and particularly, the disclosed DNA complementary palindromic patterns can be directly used to design siRNAs for RNA interference experiments, allowing more basic researchers to study ASFV infection and pathogenic mechanisms without viral infection.

Both ASFV (a dsDNA virus) and SARS-CoV (a +ssRNA virus) exhibit typical statistical characteristics of DNA complementary palindromic patterns. However, whether ASFV can produce cpsRNAs, the mechanisms of their generation, and their biological functions remain to be investigated. Based on current research, +ssRNA viruses like SARS-CoV likely generate cpsRNAs through host defense mechanism-mediated cleavage during double-stranded RNA formation. While dsDNA viruses like ASFV transcribe RNAs that can form internal double-stranded regions, whether they can generate cpsRNAs through intramolecular

cleavage remains unknown. After cpsRNA generation, whether their unique stem-loop structures can be further cleaved to produce smaller fragments (7-13 nt) and whether these possess biological functions represent important research directions. We hypothesize that such smaller RNAs may be involved in reverse transcription priming or host gene transcriptional regulation.

Acknowledgments

We thank Professor Bu Wenjun (College of Life Sciences, Nankai University), Professor Ruan Jishou (School of Mathematical Sciences, Nankai University), and Dr. Liu Chang (Medical School, Nankai University) for their long-term support. We also thank master' s students Niu Xiaoran, Ji Haishuo, and Jin Xiufeng (College of Life Sciences, Nankai University) for their participation.

References

1. Liu C, Ze Chen, Hu Y, Ji HS, Yu DS, Shen WY, Li SY, Ruan JS, Bu WJ; Gao, S (2018) Complemented palindromic small RNAs first discovered from SARS Coronavirus. Genes: in press
2. Montgomery RE (1921) On a form of swine fever occurring in British East Africa (Kenya Colony). Journal of comparative pathology and therapeutics 34:159-191
3. Galindo I, Alonso C (2017) African swine fever virus: a review. Viruses 9(5):103
4. Zhang M, Sun H, Fei Z, Zhan F, Gong X, Gao S Fastq_{clean}: An optimized pipeline to clean the Illumina sequencing data with quality control. In: Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on, 2014. IEEE, pp 44-48
5. Gao S, Ou J, Xiao K (2014) R language and Bioconductor in bioinformatics applications (Chinese Edition). Tianjin Science and Technology Translation and Publishing Co., Tianjin
6. Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D (2012) Using Tablet for visual exploration of second-generation sequencing data. Brief Bioinform: bbs012.
7. Shan G, Tian X, Sun Y, Wu Z, Cheng Z, Dong P, Zhao Q, He B, Ruan J, Bu W (2017) Two novel lncRNAs discovered in human mitochondrial DNA using PacBio full-length transcriptome data. Mitochondrion 38:41-47.
8. Chew, DSH; Choi, KP; Heidner, H; Leung, MY (2004) Palindromes in SARS and other coronaviruses. Informs J. Comput. 16:331–340.
9. Wang F; Sun Y; Ruan JS; Chen R; Chen X; Chen CJ; Kreuze JF; Fei ZJ; Zhu X; Gao S (2016) Using small RNA deep sequencing to detect human viruses. BioMed Res. Int. 2016, 2016, 2596782.

Note: Figure translations are in progress. See original paper for figures.