Precise annotation of tick mitochondrial genomes reveals multiple copy number variation of short tandem repeats and one transposon-like element

Ze Chen1, Yibo Xuan1,2, Guangcai Liang3, Xiaolong Yang1, Zhijun Yu1, Stephen C. Barker4, Samuel Kelava4, Wenjun Bu2, Jingze Liu1* and Shan Gao2,5*

Abstract

Background: In the present study, we used long-PCR amplification coupled with Next-Generation Sequencing (NGS) to obtain complete mitochondrial (mt) genomes of individual ticks and unprecedently performed precise annotation of these mt genomes. We aimed to: (1) develop a simple, cost-effective and accurate method for the study of extremely high AT-content mt genomes within an individual animal (e.g. Dermacentor silvarum) containing miniscule DNA; (2) provide a high-quality reference genome for D. silvarum with precise annotation and also for future studies of other tick mt genomes; and (3) detect and analyze mt DNA variation within an individual tick.

Results: These annotations were confirmed by the PacBio full-length transcriptome data to cover both entire strands of the mitochondrial genomes without any gaps or overlaps. Moreover, two new and important findings were reported for the first time, contributing fundamental knowledge to mt biology. The first was the discovery of a transposon-like element that may eventually reveal much about mechanisms of gene rearrangements in mt genomes. Another finding was that Copy Number Variation (CNV) of Short Tandem Repeats (STRs) account for mitochondrial sequence diversity (heterogeneity) within an individual tick, insect, mouse or human, whereas SNPs were not detected. The CNV of STRs in the protein-coding genes resulted in frameshift mutations in the proteins, which can cause deleterious effects. Mitochondria containing these deleterious STR mutations accumulate in cells and can produce deleterious proteins.

Conclusions: We proposed that the accumulation of CNV of STRs in mitochondria may cause aging or diseases. Future tests of the CNV of STRs hypothesis help to ultimately reveal the genetic basis of mitochondrial DNA variation and its consequences (e.g., aging and diseases) in animals. Our study will lead to the reconsideration of the importance of STRs and a unified study of CNV of STRs with longer and shorter repeat units (particularly polynucleotides) in both nuclear and mt genomes.

Keywords: Mitochondrial DNA, Precise annotation, Short tandem repeat, Transposon, Tick

© The Author(s). 2020 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.
Background
Annotation of mitochondrial (mt) genomes is indispensable for fundamental research in many fields, including mt biochemistry, physiology, and the molecular phylogenetics and evolution of animals. Moreover, high-resolution annotation of animal mt genomes can be used to investigate RNA processing, maturation, degradation and even the regulation of gene expression [1]. In our previous studies, two substantial contributions to the methods used to annotate mt genomes have been published. The first one was that Gao et al. constructed the first quantitative transcription map of animal mt genomes by sequencing the full-length transcriptome of the insect *Erthesina fullo* Thunberg [1] on the PacBio platform [2]. Novel findings included the 3′ polyadenylation and possible 5′ m7G caps of rRNAs [1], the polycistrionic transcripts [1], the antisense transcripts of all mt genes [1], and novel long non-coding RNAs (lncRNAs) [3]. Based on these findings, we proposed the uninterrupted transcription of mammal mt genomes [3]. In addition, we proposed that long antisense transcripts degrade quickly as transient RNAs, making them unlikely to perform specific functions [4], although all antisense transcripts are processed from two primary transcripts. The second contribution concerned the use 5′ and 3′ end small RNAs (5′ and 3′ sRNAs) [4] to annotate mt genes to a resolution of 1 bp, subsequently dubbed “precise annotation” [5]. Precise annotation of these accurate genomes led us to discover a novel 31-nt ncRNA in mammalian mt DNA [4] and that the copy numbers of tandem repeats exhibit great diversity within an *E. fullo* individual [5]. Recently, precise annotation of human, chimpanzee, rhesus macaque and mouse mt genomes has been performed to study five Conserved Sequence Blocks (CSBs) in the mt D-loop region [6]; this ultimately led to a deep understanding of the mechanisms involved in the RNA-DNA transition and even the functions of the D-loop.

In the present study, we used long-PCR amplification coupled with Next-Generation Sequencing (NGS) to obtain complete mt genomes of individual ticks and performed precise annotation of these mt genomes. Given that conventional mtDNA isolation and purification are not required in our method and in the Whole-Genome Sequencing (WGS) method, both the WGS method and our method are simple and cost-effective. However, compared to the WGS method, our method has three main advantages: (1) errors in the assembly of mt genomes caused by highly similar exogenous or nuclear sequences [i.e., Nuclear Mitochondrial DNA (NUMT)] are avoided; (2) highly similar segments (e.g., control regions 1 and 2 of *Dermacentor silvarum*) of mt genomes can be assembled separately (Results); and (3) sequence heterogeneity and DNA variation in mt genomes within an individual can be accurately determined due to the high depth of sequencing data. In the present study, we aimed to achieve the following research goals: (1) develop a simple, cost-effective and accurate method for the study of extremely high AT-content mt genomes within an individual animal (e.g. *D. silvarum*) containing miniscule DNA; (2) provide a high-quality reference genome for *D. silvarum* with precise annotation and also for future studies of other tick mt genomes; and (3) detect and analyze mt DNA variation within an individual tick.

Results
Using long-PCR and NGS to obtain complete mt genomes of individual ticks
A previous study [7] classified tick mt genomes into three types according to gene orders (Fig. 1a): (1) type I for Argasidae (soft ticks) and non-Australasian Prostriata (“other *Ixodes*”); (2) type II for Australasian Prostriata (“Australasian *Ixodes*”); and (3) type III for Metastrata (all other hard ticks). The nomenclature “other *Ixodes*” and “Australasian *Ixodes*” is from [8]. The present study focused on the genus *Dermacentor* belonging to Metastrata using ticks from four species (*D. silvarum*, *D. nuttalli*, *D. marginatus* and *D. nivipes*). The type III mt genomes of individual ticks (Fig. 1b) were obtained using long-PCR amplification coupled with NGS (Methods). All the reference genomes of tick mitochondria read in the 5′ → 3′ direction as the major coding strand (J-strand). Using specific primers (Table 1), each entire mt genome was amplified in two large segments: large segment 1 (L1) and large segment 2 (L2) or large segment 3 (L3) and large segment 4 (L4). L1 and L2 contain Control Region 1 (CR1) and Control Region 2 (CR2), respectively, whereas L3 and L4 contain tandem Repeat 1 (R1) and tandem Repeat 2 (R2), respectively (Fig. 1a).

Using ~ 4 Gbp 2 × 150 DNA-seq data for each genome, the complete mt genomes of *D. silvarum*, *D. nuttalli* and *D. marginatus* were obtained by assembling L3 and L4 separately then merging L3 and L4 (Fig. 1b). In addition, CR1 and CR2 on L4 were validated using PCR amplification coupled with Sanger sequencing, separately, as CR1 and CR2 share an identical segment (Fig. 2a). Using ~ 4 Gbp 2 × 150 DNA-seq data, the complete mt genome of *D. silvarum* was also obtained by assembling L1 and L2 separately then merging L1 and L2. Furthermore, R1 and R2 on L1 were validated using PCR amplification coupled with Sanger sequencing, separately, as the repeat units of R1 are the reverse complements of the repeat units of R2 (Fig. 3). Comparison of the *D. silvarum* mt genomes obtained by sequencing L3 and L4 with those obtained by sequencing L1 and L2 improved the accuracy of the DNA sequence. Since both R1 and R2 were longer than 150 bp, we also used ~ 4 Gbp 2 × 250 bp DNA-seq data to obtain full-
length sequences of R1 and R2 for genome polishing. In total, 12.7 Gbp DNA-seq data were generated to cover ~848,069 × (12.72 Gbp/1.5 Kbp) of the D. silvarum genome (GenBank: MN347015) which was used as a reference for precise annotation in the following studies.

Comparison of D. silvarum, D. nuttalli and D. marginatus mt genomes showed that they have the same gene order (type III) and high sequence identities (> 95%) (tandem repeats were not part of these calculations). Preliminary analysis showed two significant features in these tick mt genomes that are also possible in other ticks of Metastrongyloidea (Fig. 1a): (1) the mt genomes of D. silvarum, D. nuttalli and D. marginatus contained two tandem repeats (R1 and R2); and (2) these tick mt genomes contained multiple Short Tandem Repeats (STRs) with very short repeat units (1 or 2 bp). STRs, widely used by forensic geneticists and in studies of genealogy, are often referred to as Simple Sequence Repeats (SSRs) by plant geneticists or microsatellites by oncologists. Found widely in animal mt genomes, STRs follow a pattern in which one or more nucleotides (repeat unit) are repeated and the repeat units are directly adjacent to each other.

![Fig. 1](image)

**Table 1** PCR primers for the Dermacentor mt genomes

| Forward primer | Reverse primer | Segment | Size(bp) |
|----------------|----------------|---------|----------|
| TCAGTCATTATTACCGGATGA | GCTCAAATTCCATCTCTGC | L1 | 9580 |
| AGCTGTACTAAAGTGTTGAGG | AGGATGTGATGGATGCAAA | L2 | 8156 |
| GCTAKTGGGTCATAACCCAAA | CGACCTCGATGGTAGTGG | L3 | 7155 |
| CCAACCTGATTCACCGGCTCT | TCATGGCGGTAAGATGACTGA | L4 | 9187 |
| TGCTGCTGGCACAAATTTGAC | CAAGATGCCCTAAATTCAGCGCA | CR1 | 483 |
| GGAGCTATACCATCCTCAACAGC | TTGAGGAATGAACCCAATAGC | CR2 | 645 |
| TGACTCTGATTGATGCTCCT | CCGGGCTATCTCATATGACC | R2 | 3616 |
| CTATTCCCGCATAGAATATGCCGTG | CAAGCTATGGCACCCTTCATACAT | R1 | 570 |

These primers were designed to amplify large segments (L1, L2, L3 and L4) and short segments (CR1, CR2, R1 and R2) in the mt genomes of the genus Dermacentor. Their PCR reaction conditions can be seen in the Methods. Based on the results using 100 individual ticks from four species, the primers for L3 and L4 were optimized to amplify more species of the genus Dermacentor than those of L1 and L2. The R2 segment spanned tRNA^Asp^, tRNA^Glu^, tRNA^Glu^, ND1, tRNA^Val^, 16S rRNA, tRNA^Val^, 125 rRNA and CR1. The R1 segment spanned tRNA^Asp^, tRNA^Glu^, R1, tRNA^Glu^ and NDS. The segment sizes were estimated using the D. silvarum mt genome (GenBank: MN347015).
each other, allowing for very rare Single Nucleotide Polymorphisms (SNPs) in the repeat units. The minimum length of the repeat units of STRs is obviously 1 bp; we call type of STR a polynucleotide. PolyAs and polyTs occur frequently in tick and insect mt genomes; indeed, they contribute substantially to the high AT content of many of these mt genomes. Polynucleotides and tandem repeats R1 and R2 had the same pattern of variation in the D. silvarum mt genome (below). This suggested that a unified study should be performed on the CNV of STRs with longer and shorter repeat units, particularly polynucleotides that were usually overlooked in previous studies. To describe a tandem repeat, we use the repeat unit and its copy number. STRs can be classified by their repeat unit length (m) and copy number (n), thus briefly noted as \( m \times n \) STR. For example, the STR ATATATATAT is noted as \([AT]_5\) and classified as \(2 \times 5\) STR. In this way, a polynucleotide is classified as \(1 \times n\) STR.

**Precise annotation of the Dermacentor silvarum mt genome**

Our D. silvarum mt genome shares a sequence identity of 97.47% with the publicly available D. silvarum mt genome NC_026552.1 in the NCBI RefSeq database. We performed precise annotation of the complete D. silvarum mt genome (Table 2) using sRNA-seq data and confirmed these annotations using the PacBio full-length transcriptome data (Methods). Although most of the new annotations were consistent with those of NC_026552.1, we corrected many errors in NC_026552.1, particularly in tRNAs, rRNAs, CR1, CR2, R1 and R2.

D. silvarum transcribes both entire strands of its mt genome to produce primary transcripts covering CR1 and CR2, predicted to be non-coding and non-transcriptional regions in a previous study [7]. CR1 with a length of 309 bp and CR2 with a length of 307 bp shared a 263-bp identical segment (Fig. 2a). CR1 and R1 were annotated as full-length RNAs cleaved from the minor coding strand (N-strand) primary transcript, whereas tRNA\(^{\text{Ala}}\), tRNA\(^{\text{Glu}}\), tRNA\(^{\text{Tyr}}\) and tRNA\(^{\text{Phe}}\) had unstable T-arms (indicated in black box).

Using precise annotations, we obtained two new findings about the D. silvarum mt tRNAs. The first involved six mt tRNA genes, from which atypical tRNAs with no
D-arm or an unstable T-arm were inferred [9]. One of tRNA\textsuperscript{A\textsubscript{5}}s (mtDNA: 5713:5769) and tRNA\textsuperscript{Cys} (Fig. 2b) had no D-arms, whereas tRNA\textsuperscript{A\textsubscript{5}} (Fig. 2b), tRNA\textsuperscript{Glu}, tRNA\textsuperscript{Tyr} and tRNA\textsuperscript{Phe} had unstable T-arms. Another new finding was that the intergenic regions between tick mt tRNA genes are longer than those in mammals except a novel 31-nt ncRNA [4], which was generated in the gene order rearrangement of mammalian mt tRNA genes. Although these intergenic regions in ticks were cleaved between their neighbouring tRNAs to form small RNAs (sRNAs) shorter than 10 bp, they are not likely to have biological functions, in our view. One typical example of a sRNA was A\textsuperscript{U\textsubscript{7}}, between tRNA\textsuperscript{Cys} and tRNA\textsuperscript{Met} (Fig. 2b). Based on these two findings, we found that 1 × n STRs involved both intergenic regions (e.g., A\textsuperscript{U\textsubscript{7}}), and atypical mt tRNAs (e.g., A\textsubscript{5} in tRNA\textsuperscript{Cys}). Comparison of tRNA\textsuperscript{A\textsubscript{5}} and tRNA\textsuperscript{Cys} suggested that tRNA\textsuperscript{Cys} (Fig. 2b) with no D-arm had an A\textsubscript{5} insertion that formed a large loop. Given that the tRNA\textsuperscript{Cys} DNA sequence had too little evolutionary conservation to allow for a STR insertion, it proved a long-standing hypothesis that atypical tRNAs do not have biological functions.

R1 and R2 (Fig. 3) were predicted to be two non-coding and non-transcriptional regions in the previous study [7]. In the present study, however, they were proven to be transcribed on two strands. The repeat units in R1 were reverse complements of those in R2 (Fig. 3b). Our DNA-seq data showed that the copy numbers of R1 and R2 exhibited great diversity within an individual, which confirmed a finding from our previous study of the E. fullo mt genome [5]. Since repeat units in R1 and R2 were reverse complements, we used PCR amplification (Table 1) coupled with Sanger sequencing to further investigate R1 sequences in more than 100 individual ticks from four species (D. silvarum, D. nuttalli, D. marginatus and D. niveus) and obtained the following results: (1) for each individual tick, the R1 sequence obtained using Sanger sequencing is actually a consensus sequence of a large number of heterogeneous sequences; (2) copy numbers were distributed between 2 and 5 for all studied repeat units, with one partial repeat unit counted as 1; (3) in total, three types of repeat units (type 1, 2 and 3) of R1 were identified; (4) in general, R1 sequences from individual ticks of one species comprised repeat units of one type and R1 sequences from individual ticks of the same species from different places could have different copy numbers; and (5) of the four species of ticks we studied, D. nuttalli and D. niveus had a R1 which was composed of type 3 units, whereas D. silvarum had a R1 which was composed of type 2 or type 3 units. As for D. marginatus, most of the R1s were composed of type 3 units; however, a few had R1s composed of the types 1 and 2 hybrid units, noted as [R\textsubscript{34}l−R\textsubscript{34}m−R\textsubscript{34}n], where l, m and n represent the copy numbers. The discovery of these "hybrid" units suggested to us that mt DNA recombination may occur within an individual tick,
| Gene          | Strand | Start | End  | Length |
|--------------|--------|-------|------|--------|
| tRNA-Met     | (+)    | 1     | 67   | 67     |
| ND2          | (+)    | 68    | 1028 | 961    |
| tRNA-Trp     | (+)    | 1029  | 1091 | 63     |
| tRNA-TyrAS   | (+)    | 1092  | 1149 | 58     |
| COI          | (+)    | 1150  | 2686 | 1537   |
| COII         | (+)    | 2687  | 3359 | 673    |
| tRNA-Lys     | (+)    | 3360  | 3429 | 70     |
| tRNA-Asp     | (+)    | 3430  | 3491 | 62     |
| ATP8/6       | (+)    | 3492  | 4326 | 835    |
| COIIII       | (+)    | 4327  | 5104 | 778    |
| tRNA-Gly     | (+)    | 5105  | 5171 | 67     |
| ND3          | (+)    | 5172  | 5511 | 340    |
| tRNA-Ala     | (+)    | 5512  | 5576 | 65     |
| Intergenic   | (+)    | 5577  | 5578 | 2      |
| tRNA-Arg     | (+)    | 5579  | 5640 | 62     |
| tRNA-Asn     | (+)    | 5641  | 5708 | 68     |
| Intergenic   | (+)    | 5709  | 5712 | 4      |
| tRNA-Ser     | (+)    | 5713  | 5769 | 57     |
| tRNA-Glu     | (+)    | 5770  | 5834 | 65     |
| R2           | *      | 5825  | 5987 | 163    |
| HAS1         | (+)    | 5835  | 9258 | 3424   |
| tRNA-Ile     | (+)    | 9259  | 9322 | 64     |
| HAS2         | (+)    | 9323  | 12,930 | 3608 |
| tRNA-Thr     | (+)    | 12,931 | 12,991 | 61    |
| tRNA-ProAS/ND6 | (+)    | 12,992 | 13,503 | 512   |
| Cytb         | (+)    | 13,504 | 14,585 | 1082  |
| tRNA-Ser     | (+)    | 14,586 | 14,652 | 67    |
| tRNA-LeuAS/CR2 | (+)    | 14,653 | 15,023 | 371   |
| CR2          | *      | 14,717 | 15,023 | 307   |
| tRNA-Cys     | (+)    | 15,024 | 15,086 | 63    |
| Intergenic   | (+)    | 15,087 | 15,094 | 8     |
| tRNA-Tyr     | (−)    | 1090  | 1151 | 62     |
| LAS1         | (−)    | 1152  | 5984 | 4833   |
| ND1          | (−)    | 5985  | 6903 | 919    |
| tRNA-Leu     | (−)    | 6904  | 6964 | 61     |
| 16S rRNA     | (−)    | 6965  | 8185 | 1221   |
| tRNA-Val     | (−)    | 8186  | 8247 | 62     |
| 12S rRNA     | (−)    | 8248  | 8949 | 702    |
| CR1          | (−)    | 8950  | 9258 | 309    |
| tRNA-IleAS   | (−)    | 9259  | 9323 | 65     |
| tRNA-Gln     | (−)    | 9324  | 9390 | 67     |
| R1           | (−)    | 9391  | 9559 | 169    |
| tRNA-Phe     | (−)    | 9560  | 9620 | 61     |
| NDS          | (−)    | 9621  | 11,275 | 1655  |
resulting in the insertion of $[R_{87}]_m$ into $[R_{87}]_n$. This confirmed our proposal of DNA-recombination events in a previous study of the *E. f. silvarum* mt genome [5]. In that previous study, the insertion of segments A and B into STR $[R_{87}]_1$ resulted in $[R_{87}]_1$-$[R_{87}]_n$.

### Discovery of a transposon-like element

In a previous study, the repeat unit was conceived as the “tick box”—a degenerate 17-bp sequence motif that may be involved in the 3′ formation of ND1 and tRNA$^{\text{Glu}}$ transcripts in all major tick lineages [7, 10, 11]. A large translocated segment (LT1) spanning from ND1 to tRNA$^{\text{Gln}}$ was first reported in 1998 [12, 13] and the presence of the “tick box” motif at both ends of this LT1 indicated its involvement in recombination events that are responsible for known Metastriata ticks [12–14]. Metastriata genome rearrangements have been found in all Metastriata ticks studied [8, 10–18] (Fig. 1a). In the present study, LT1 was corrected to span R2, ND1, tRNA$^{\text{Leu}}$, 16S rRNA, tRNA$^{\text{Val}}$, 12S rRNA, CR1, tRNA$^{\text{Ile}}$, tRNA$^{\text{Gln}}$ and R1 (Fig. 3a) in the reference genome using precise annotations. Given that nearly half of the human genome is various types of transposable elements that contain repetitive DNA sequences [19], we hypothesized that LT1 is a transposon, with R1 and R2 as invert repeats (IRs) and genes from ND1 to tRNA$^{\text{Gln}}$ as insert sequences (ISs). To test our hypothesis, we sought structural variation (Methods) in the *D. silvarum* mt genome to determine the occurrence of LT1 translocation events. The results proved the occurrence of LT1 inversions within an individual tick (Fig. 3a).

Since LT1 inversions were rare, 4.1 Gbp DNA-seq data were generated to cover ~427,247× (4.09 Gbp/9.58 Kbp) of L1 in the *D. silvarum* mt genome to detect the LT1 inversions. As the dominant copy number was five for both R1 and R2, we used 34×5 STR to represent R1 and R2 in the *D. silvarum* mt genome (Fig. 3c). Thus, R1 and R2 in *D. silvarum* are ~170-bp long (34 × 5), which is longer than the reads in the 2 × 150 bp DNA-seq data. We had to sequence the same library using 2 × 250 bp sequencing to validate the reference genome and the LT1 inversion (Methods). The substantial diversity in R1 and R2 copy numbers within an individual tick rendered great diversity in LT1. However, we did not obtain full-length sequences of LT1 due to sequence-length limitations in the DNA-seq data. Therefore, we were unable to determine whether R1 and R2 had the same copy numbers within one LT1.

### Copy number variation of STRs in the mt genomes within an individual animal

By mapping DNA-seq data to the *D. silvarum* mt genome, variation detection (Methods) was performed to report two types of DNA variation—SNPs and small insertions/deletions (InDels). Almost all the detected DNA variation within a *D. silvarum* tick was Copy Number Variation (CNV) of STRs caused by InDels of one or more entire repeat units, whereas SNPs were not detected. We defined the STR position as the genomic position of the first nucleotide of the reference STR. For example, $[G]_9$ was designated as the reference STR at position 1810, because it occurred most frequently in mtDNAs within one individual tick (Table 3); the alternative allele of $[G]_9$ included $[G]_{10}$, $[G]_7$, $[G]_9$ and $[G]_{10}$. Importantly, it was found that almost all of the STRs had multiple variants, particularly those with copy numbers greater than 5. The detection of CNV of STRs was reliable, based on the following reasons: (1) PCR amplification and deep DNA sequencing produces a high signal-to-noise ratio in the detection of DNA variation; (2) the Illumina sequencer generates very rare InDel errors, i.e. few per million bases [21]; (3) it is impossible for sequencing or alignment errors to result in 2-bp InDels in 2×n STR (e.g., $[TA]_9$); (4) the alternative allele ratios (Methods) at some positions were significantly higher and the highest ratio reached was ~32% at

### Table 2 Precise annotations of the *Dermacentor silvarum* mt genome (Continued)

| Gene       | Strand | Start | End   | Length |
|------------|--------|-------|-------|--------|
| tRNA-His   | (-)    | 11,276| 11,341| 66     |
| ND4/4L     | (-)    | 11,342| 12,928| 1587   |
| tRNA-ThrAS | (-)    | 12,929| 12,991| 63     |
| tRNA-Pro   | (-)    | 12,992| 13,055| 64     |
| LAS2       | (-)    | 13,056| 14,654| 1599   |
| tRNA-Leu   | (-)    | 14,655| 14,716| 62     |
| LAS3       | (-)    | 14,717| 1089  | 1467   |

This reference sequence is available at the NCBI GenBank database under the accession number MN347015. J(+), J(−) and N(−) represent the major and minor coding strands of the mt genome, respectively. Control Region 1 (CR1) and tandem Repeat 1 (R1) were annotated as full-length RNAs cleaved from the minor coding strand (N-strand) primary transcript, whereas CR2 and R2 were annotated as DNA regions (>). The “AS” suffix represents antisense. H-strand Antisense Segment 1 (HAS1) represents R2/ND1AS/tRNA$^{\text{Gln}}$/tRNA$^{\text{Phe}}$/tRNA$^{\text{His}}$/ND5AS/tRNA$^{\text{Glu}}$/CR1. HAS2 represents tRNA$^{\text{Gln}}$/ND6AS/CytbAS/tRNA$^{\text{Ser}}$. LAS1 represents COIAS/COIIAS/tRNA$^{\text{Tyr}}$/tRNA$^{\text{His}}$/tRNA$^{\text{Gly}}$/CR1. LAS2 represents COIIAS/COIIAS/tRNA$^{\text{Asp}}$/tRNA$^{\text{Ser}}$/tRNA$^{\text{Ser}}$/tRNA$^{\text{Ser}}$/tRNA$^{\text{Ser}}$/tRNA$^{\text{Ser}}$/tRNA$^{\text{Ser}}$/tRNA$^{\text{Ser}}$/tRNA$^{\text{Ser}}$.
position 3441 (Table 3); (5) the same CNV of STRs caused diversity within and between individual ticks, e.g., [TA]9 in our genome (Table 3) and [TA]5 in NC_0265521; (6) CNV of STRs was detected in other animal species, including E. fullo [5], mouse; and (7) Almost all the detected DNA variation within a single human cell were CNV of STRs, whereas SNPs were not detected using a colon cancer single cell RNA-seq (scRNA-seq) dataset (SRA: SRP113436).

Using a very strict parameter (Methods), we selected 20 STR positions in the D. silvarum mt genome for further study (Table 3). STRs at 18 positions were 1 × n STR and those at the other 2 positions were 2 × n STR. Almost all of the STRs were composed of A or T, except [G]9 at position 1810. All of the reference STRs and their variants had copy numbers greater than 5. Among all of the variants, 1 × n STR occurred much more frequently than n × n STR (n > 1). Thirteen STR positions in the protein-coding genes had identical sequences between individuals, exhibiting a high degree of evolutionary conservation; however, the other seven STR positions in the tRNA and rRNA genes exhibited variation. For example, [A]10, [TA]9, [T]9 and [T]10 at positions 5524, 7324, 8042 and 8243 in our mt genome changed to [A]8, [TA]5, [T]8 and [T]6 at positions 5524, 7324, 8042 and 8243 in NC_0265521 (Table 3). The CNV of STRs in the protein-coding genes resulted in frameshift mutations in the proteins, which may be deleterious [22]. For example, the COI gene has a 1524-bp Coding Sequence (CDS) for a 508-aa protein. The variant [G]9 at position 1810 (Table 3) results in a 765-bp CDS for a 255-aa truncated COI protein. This finding inspired us to investigate if animal cells have mechanisms to remove mitochondria containing deleterious mutations or inhibit the expression of the deleterious variants, as they can cause loss of function or diseases. We did not, however, obtain deep RNA-seq data for D. silvarum. We had to compare the STR variants in the E. fullo mt genome [5] at the DNA level and RNA levels, using the DNA-seq and RNA-seq data (SRA: SRP174926). However, we found no significant differences between them. This suggested that deleterious STR mutations can irreversibly change the proteins made from their mRNAs and that mitochondria containing deleterious mutations may accumulate in cells.

**Discussion**

SNPs, accepted as the most common type of genetic variation, play a dominant role in studies in almost all
biological fields; however, research of STRs has been limited to specialized fields. In particular, the use of SNPs is becoming dominant in the studies of mitochondria, e.g., mt heterogeneity. After further analysis of tick, insect, mouse and human DNA-seq and RNA-seq data, we found that a large number of STR variants had been missed in variation detection, as most research has focused only on the detection of SNPs using WGS data. Although SNP platforms are higher throughput and more cost-effective for genome scans, STRs remain highly informative measures of DNA variation for linkage and association studies. Using PCR coupled with very deep DNA-seq data, we were able to allow three gaps in genome alignment and still ensure the detection accuracy of STRs. Based on these results, we proved that WGS and RNA-seq data can also be used to detect CNV of STRs with careful adjustment of parameters (Methods). Future studies are necessary on the origins, mutation rates and effects of CNV of STRs in, but not limited to, animal mt genomes.

Both 1 × n STRs and 34 × n STRs had the same variation pattern; 34 × n STRs can be produced by DNA recombination, however, the cause of 1 × n STRs remains unclear. One possible cause of 1 × n STRs is replication slippage [23]. According to the “replication slippage” theory, DNA polymerase causes mismatches between DNA strands while being replicated during meiosis. The DNA polymerase can then slip while moving along the template strand and continue on the wrong nucleotide. Another cause might be point mutations, based on a study comparing human and primate genomes [24]. Both replication slippage and point mutations cannot be used to explain the occurrence of m × n STRs (m = 2, 3, 4, 5...). As both 1 × n STRs and 34 × n STRs had the same variation pattern, they are probable to have the same causes.

Unique DNA sequences in a genome have a very low variation/mutation rate (approximately 10^{-9} nt per generation), whereas the mutation rates in STRs are several orders of magnitude higher, ranging from 10^{-6} to 10^{-2} nt per generation for each locus [25]. Direct estimates of STR mutation rates have been made in numerous organisms from insects to humans, e.g., Schistocerca gregaria 2.1 × 10^{-4} [26]. However, most research focuses on SNPs, rather than CNV of STRs in mt genomes. In the present study, we only detected CNV of STRs in mt genomes within an individual animal and found that the alternative allele ratio was distributed from less than 0.01% to ~ 33%. This suggested that CNV of STRs occurrences varied significantly along the genomes such that mutations concentrated at certain positions (e.g., 20 STR positions for the D. silvarum mt genome). Future studies must be performed to estimate STR mutation rates of mitochondria to test if they have correlations with life expectancy of different animals, using individuals at different developmental stages.

It is well accepted that many STRs are located in non-coding DNA and are biologically silent, while others are located in regulatory or even coding DNA. STRs located in regulatory, intron and transposon regions are beyond the scope of the present study. However, our studies showed there were no significant differences between the alternative allele ratios of STR positions in protein-coding gene regions and those in tRNA and rRNA gene regions in tick mt genomes. Particularly, mitochondria containing deleterious mutations can accumulate in cells and deleterious STR mutations irreversibly change the proteins made from their mRNAs. This suggested that deleterious STR mutations in mitochondria cause aging and diseases, also known as Tandem Repeat Disorders (TRDs). Huntington’s disease, as one of the famous TRDs, occurs in the context of expanded glutamine [CAG]_n repeats. Several other human diseases have also been linked to CNV of mt STRs. One famous example—breast cancer (BC)—has been linked to D_310 variation [27]. The telomeres at the ends of the chromosomes consist of [TTAGGG]_n in vertebrates. Thus, both telomeres and mitochondria are linked to ageing/senescence by CNV of STRs.

**Conclusion**

In the present study, we used long-PCR amplification coupled with NGS to obtain complete mt genomes of individual ticks and performed precise annotation of these genomes. The discovery of a transposon-like element shed light on the mechanisms of mt gene order rearrangement and genomic structural variation, especially with additional data from more tick species. The second finding may pave the way to an eventual understanding of the mechanisms of mt DNA variation. The comparison between interindividual and intraindividual variation showed that STRs with shorter repeat units (e.g., 1 × n STRs) and STRs with longer repeat units (e.g., 34 × n STRs) had the same variation pattern. This finding will encourage reconsideration of the importance of STRs as well as a unified study of CNV of STRs with shorter and longer repeat units in both nuclear and mt genomes.

**Methods**

Individual ticks from four species (D. silvarum, D. nuttalli, D. marginatus and D. niveus) of the genus Dermacentor were collected from different places in China and were identified using a stereoscopic microscope according to [28]. Total DNA was isolated from individual ticks using QIAamp DNA Mini Kit (Qiagen, Germany), following its protocol. Long-PCR amplification of each entire mt genome in L1, L2, L3 and L4 (Fig. 1a) was performed using Long PCR Mix (Sino-Novel, China)
and the PCR reaction mixture was also used to construct one ~350-bp-size library and sequenced using 2 × 150 bp paired-end strategy on the Illumina HiSeq X Ten platform. The PCR-amplified L1 of D. silvarum was submitted to the NCBI SRA database under the accession number SRP178347. The complete mt genome sequence of D. silvarum is available at the NCBI GenBank database under the accession number MN347015. However, precise annotations of the D. silvarum mt genome need be seen in Table 2, as the NCBI GenBank database is not able to accept this new annotation format. The PCR-amplified short segments (Table 1) were sequenced using Sanger sequencing. Using the software BWA v0.5.7, DNA-seq reads were aligned to the D. silvarum mt genome with parameters (aln -n 3 -o 1 -e 2 -l 15 -k 1) to detect DNA variation. CNV of STRs (Table 3) was detected by single-end alignment using the 2 × 150 bp DNA-seq data of L1 and L2, and then, validated using the 2 × 150 bp DNA-seq data of L3 and L4 and 2 × 250 bp DNA-seq data of L1. The alternative allele ratio was calculated by the depth of all alternative alleles divided by the total depth on the genomic position. The 20 STR positions were selected using a strict parameter, which required the alternative allele ratio above 1%.

Total RNA was isolated from ticks to construct sRNA-seq libraries, following the protocol [29] and the cleaning and quality control of sRNA-seq and the PacBio full-length transcriptome data was performed using the pipeline Fastq_clean [30]. Using the software bowtie v0.12.7, sRNA-seq reads were aligned to the D. silvarum mt genome with one mismatch for precise annotation. To confirm the precise annotation of the D. silvarum mt genome, the PacBio full-length transcriptome data was used to follow the same procedure as [5]. The structural variation detection was performed following the protocol as [31]. Statistics computing and graphics were conducted using the software R v2.15.3 the Bioconductor packages [32].

Abbreviations
NGS: Next-Generation Sequencing; mt: mitochondrial; CNV: Copy Number Variation; STR: Short Tandem Repeat; SNP: Single Nucleotide Polymorphism; NCB: National Center for Biotechnology Information; SRA: Sequence Read Archive; lncRNAs: long non-coding RNAs; sRNA: small RNA; mt: mitochondrial; ncRNA: Non-coding RNA; CR: Control Region; J-strand: major coding strand; DNA-seq: DNA sequencing; SSR: Simple Sequence Repeat; R: tandem repeat; sRNA-seq: small RNA sequencing; N-strand: minor coding strand; LT: large translocated segment; R: insert repeat; IS: insert sequence; InDel: insertion/deletion; sRNA-seq: single cell RNA-seq; CDS: Coding Sequence; TRD: Tandem Repeat Disorder; BC: breast cancer

Authors’ contributions
ZC and SG conceived this project. SG and JL supervised this project. SG, ZC and SG, ZC, SG, ZC and SG contributed equally from Professor Jishou Ruan from School of Mathematical Sciences, Nankai University. We also appreciate the help equally from the people listed below. They are Professor Guoqing Liu, Dawei Huang, Yanqiang Liu, Associate Professor Bingjun He, Qiang Zhao, the graduate student Xiufeng Jin, Hashuo Ji from College of Life Sciences, Nankai University. We would like to thank Editage (www.editage.cn) for English language editing.

Acknowledgments
We appreciate the help equally from the people listed below. They are Professor Guoqing Liu, Dawei Huang, Yanqiang Liu, Associate Professor Bingjun He, Qiang Zhao, the graduate student Xiufeng Jin, Hashuo Ji from College of Life Sciences, Nankai University. We would like to thank Editage (www.editage.cn) for English language editing.

Funding
This work was supported by National Natural Science Foundation of China (31470167) to Ze Chen, Tianjin Key Research and Development Program of China (19FYCS00500) to Shan Gao and Science Foundation of Hebei Normal University (L2020B17) to Ze Chen. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The complete mitochondrial genome sequence of Dermacentor silvarum is available at the NCBI GenBank database under the accession number MN347015 and the raw data is available at the NCBI SRA database under the accession number SRP178347.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no financial and non-financial competing interest.
Scherer, S. A short guide to the human genome. New York: Cold Spring Harbor Laboratory Press; 2008.
10. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
11. Schirmer M, D’Amore R, Iaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016;17(1):1–15.
12. Gao S, Zhang N, Duan GY, Yang Z, Ruan JS, Zhang T. Prediction of functional changes associated with single-point protein mutations using support vector machines (SVMs). Hum Mutat. 2009;30(8):1161–6.
13. Tautz D, Schlötterer C. Simple sequences. Curr Opin Genet Dev. 1994(4):832–7.
14. Amos W. Mutation biases and mutation rate variation around very short human microsatellites revealed by human–chimpanzee–orangutan genomic sequence alignments. J Mol Evol. 2010;71(3):192–201.
15. Fan H, Chu JY. A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics. 2007;5(1):7–14.
16. Chaupis MP, Plantamp C, Steffl R, Blodin L, Piot C. Microsatellite evolutionary rate and pattern in Schistocerca gregaria inferred from direct observation of germline mutations. Mol Ecol. 2015;24(26):6107–19.
17. Alhomidi MA, Vedicherla B, Movva S, Rao PK, Ahuja YR, Hasan Q. Mitochondrial D310 instability in Asian Indian breast cancer patients. Tumor Biol. 2013;34(4):2427–32.
18. Chen Z, Yang X. Systematics and taxonomy of Ixodida. Beijing: Science Press; 2019.
19. Chen Z, Sun Y, Yang X, Wu Z, Guo K, Niu X, et al. Two featured series of rRNA-derived RNA fragments (rRFs) constitute a novel class of small RNAs. PLoS One. 2017;12(4):1–9.
20. Zhang M, Zhan F, Sun H, Gong X, Fei Z, Gao S. Fastq_clean: An optimized pipeline to clean the Illumina sequencing data with quality control. In: Proceedings of 2014: IEEE. 2014: 44–48.
21. Zhang F, Xu T, Mao L, Yan S, Chen X, Wu Z, et al. Genome-wide analysis of Dongxiang wild rice (Oryza rufipogon Griff.) to investigate lost/acquired genes during rice domestication. BMC Biol. 2016;16(1):1–11.
22. Gao S, Ou J, Xiao K. R language and bioconductor in bioinformatics applications(Chinese edition). Tianjin: Tianjin Sci Technol Translation Publishing Ltd.; 2014.