Zhang et al. BMC Genomics (2020) 21:563
https://doi.org/10.1186/s12864-020-06949-5

BMC Genomics

RESEARCH ARTICLE Open Access

Relatively semi-conservative replication and a folded slippage model for short tandem repeats

Hongxi Zhang†, Douyue Li†, Xiangyan Zhao†, Saichao Pan†, Xiaolong Wu, Shan Peng, Hanrou Huang, Ruixue Shi and Zhongyang Tan*

Abstract

Background: The ubiquitous presence of short tandem repeats (STRs) in virtually all genomes implicates their functional relevance, while a widely-accepted definition of STR is yet to be established. Previous studies majorly focus on relatively longer STRs, while shorter repeats were generally excluded. Herein, we have adopted a more generous criteria to define shorter repeats, which has led to the definition of a much larger number of STRs that lack prior analysis. Using this definition, we analyzed the short repeats in 55 randomly selected segments in 55 randomly selected genomic sequences from a fairly wide range of species covering animals, plants, fungi, protozoa, bacteria, archaea and viruses.

Results: Our analysis reveals a high percentage of short repeats in all 55 randomly selected segments, indicating that the universal presence of high-content short repeats could be a common characteristic of genomes across all biological kingdoms. Therefore, it is reasonable to assume a mechanism for continuous production of repeats that can make the replicating process relatively semi-conservative. We have proposed a folded replication slippage model that considers the geometric space of nucleotides and hydrogen bond stability to explain the mechanism more explicitly, with improving the existing straight-line slippage model. The folded slippage model can explain the expansion and contraction of mono- to hexa- nucleotide repeats with proper folding angles. Analysis of external forces in the folding template strands also suggests that expansion exists more commonly than contraction in the short tandem repeats.

Conclusion: The folded replication slippage model provides a reasonable explanation for the continuous occurrences of simple sequence repeats in genomes. This model also contributes to the explanation of STR-to-genome evolution and is an alternative model that complements semi-conservative replication.

Background

Short tandem repeats (STRs), also referred as simple sequence repeats (SSRs), have attracted increasingly great interests in recent decades [1–7], and have been widely analyzed in the sequences of eukaryotic, prokaryotic and also viral genomes [2, 5, 6, 8]. STRs are the most variable genomic sequences, which tend to appear frequent variations in repeat-unit number instead of nucleotide substitution, and they may be a critical power accelerate the genomic evolution [5, 9], have roles associate with the host-adaptation and pathogenicity [9, 10], be relevant with the expression of genes and activity of promoters [4, 11], have relationship with many genetic diseases [12–14], and be observed with microsatellite instability (MSI) in many type of cancers [15–18].

* Correspondence: zhongyangtan@yeah.net
†Hongxi Zhang, Douyue Li, Xiangyan Zhao and Saichao Pan contributed equally to this work.
Bioinformatics Center, College of Biology, Hunan University, Changsha 410082, China

Though STRs have been comprehensively researched, there is actually no precise definition or wide-convinced standard for the extraction of STRs all the time, which is usually based on setting the minimum numbers of the iterations for the mononucleotide to hexanucleotide repeats based on empirical criterion [2, 3, 5, 9, 19, 20]. Majority of previous studies showed more interests into the relatively longer repetitive sequences [21–23], and most studies usually used the thresholds of 6, 3, 3, 3, 3, 3 for extracting mono- to hexanucleotide repeats [24–27], while the very short repeat-motifs with smaller iterations were almost excluded, causing the neglect of their important significance [28–31]. In this work, the selected STRs were extensively extracted with a wider extracting standard for extensive repeat-motif grabbing to investigate the essential occurrences of STR.

It is widely accepted that DNA slippage is thought to be the primary mechanism for driving STR expansion or contraction, however, slippage involves DNA polymerase pausing, dissociation and re-association [5, 32, 33], which may help to understand the expansion and contraction of long repeat sequences; it seems difficult to explain the remain of high percentage of short repeat sequences, and therefore, it is necessary to improve the slippage model more explicit to explain the generation of large amounts of short repeat sequences [34–37]. It was suggested that the STRs are most possibly born in the process of replication [5]; replication is considered to be exactly semi-conservative with that the number of nucleotides in replicating chain is be precisely equal to that in template chain, and the replicating DNA molecule was shown as a straight molecule in vitro [38, 39]. Though it is well known that the DNA molecule is highly bent and packed in a super helix state within the nucleus, the replicating DNA molecule was also believed to be dragged to a straight molecule by the polymerase complex in vivo [40–43]. But there are a lot of environmental elements inside the nucleus which may disturb the polymerase complex, and these disturbances sometimes may affect the dragged straight DNA molecule returning to some extent of bent. The bent replicating DNA molecule is possibly related to the polymerase slippage for the occurrence of short STRs. Here, we calculated the bent replicating DNA molecule with strictly considering the geometric space, the relationship between the phosphodiester bond and hydrogen bond, and also the stability of paired nucleotides; and proposed a folded replication slippage model for explaining repeats occurrence, which seems more reasonable to explain the remaining of high percentage short repeats in genomes, and also to explain the frequent STR expansion and contraction. This work may also put forward some constructive suggestions for complementing the theory of semi-conservative replication.

Here, we calculated the bent replicating DNA molecule with strictly considering the geometric space, the relationship between the phosphodiester bond and hydrogen bond, and also the stability of paired nucleotides; and proposed a folded replication slippage model for explaining repeats occurrence, which seems more reasonable to explain the remaining of high percentage short repeats in genomes, and also to explain the frequent STR expansion and contraction. This work may also put forward some constructive suggestions for complementing the theory of semi-conservative replication.

Results

Genomes tend to produce short repeats

We analyzed 55 randomly-selected sequence segments covering animal, plant, fungus, protist, bacteria, archaea and virus (Table S1). The STRs were extracted from all these sequence segments using a threshold with minimum length of 3 base pairs or nucleotides. Though 2 iteration of di-, tri-, tetra-, penta- and hexa- nucleotide repeat sequence are usually ignored in most previous studies [2, 5, 28, 29, 31], we found that the abundance of such repetitive sequences cannot be justified by the theory of random occurrence. Moreover, iteration of 3 to 5 of mononucleotide repeats also cannot be justified as random sequences. Therefore, we adapted a much more generous set of thresholds for the definition of short STRs as 3, 2, 2, 2, 2, 2 for mono-, di-, tri-, tetra-, penta-, hexa- nucleotide repeats, respectively. Aiming to analyze unexplored shorter simple repeats. The resulting sequences from this generous set of thresholds were compared with those from another two set of thresholds. As a control experiment to rule out unintentional amplification of noise, we generated mimic sequences with the same size and nucleotide composition to the corresponding 55 reported sequences.

The analyzed data showed that the reported sequence segments comprise 36.4 to 60.0% STRs under the new threshold, with an average of 44.4% (Fig. 1a, Table S1), while comparative analysis using existing standards yielded only an average of 18.8 and 5.0% STR contents on the same dataset. Since all these segments were randomly selected from their genomes, our results suggested that the high content of short STRs is a general feature of all organism genomes after long time evolution, and that the few formerly well-studied repeats may only stand for the proverbial tip of the iceberg [2, 3, 5, 6, 8]. The null hypothesis test demonstrated that the percentages of STRs in the generated segments are all lower than those in the reported segments, indicating that the high percentages of short STRs preserved valuable signals in all reported segments.

Though the evolutionary mechanism of nucleotide sequences is still hotly debated by evolutionist, it is widely

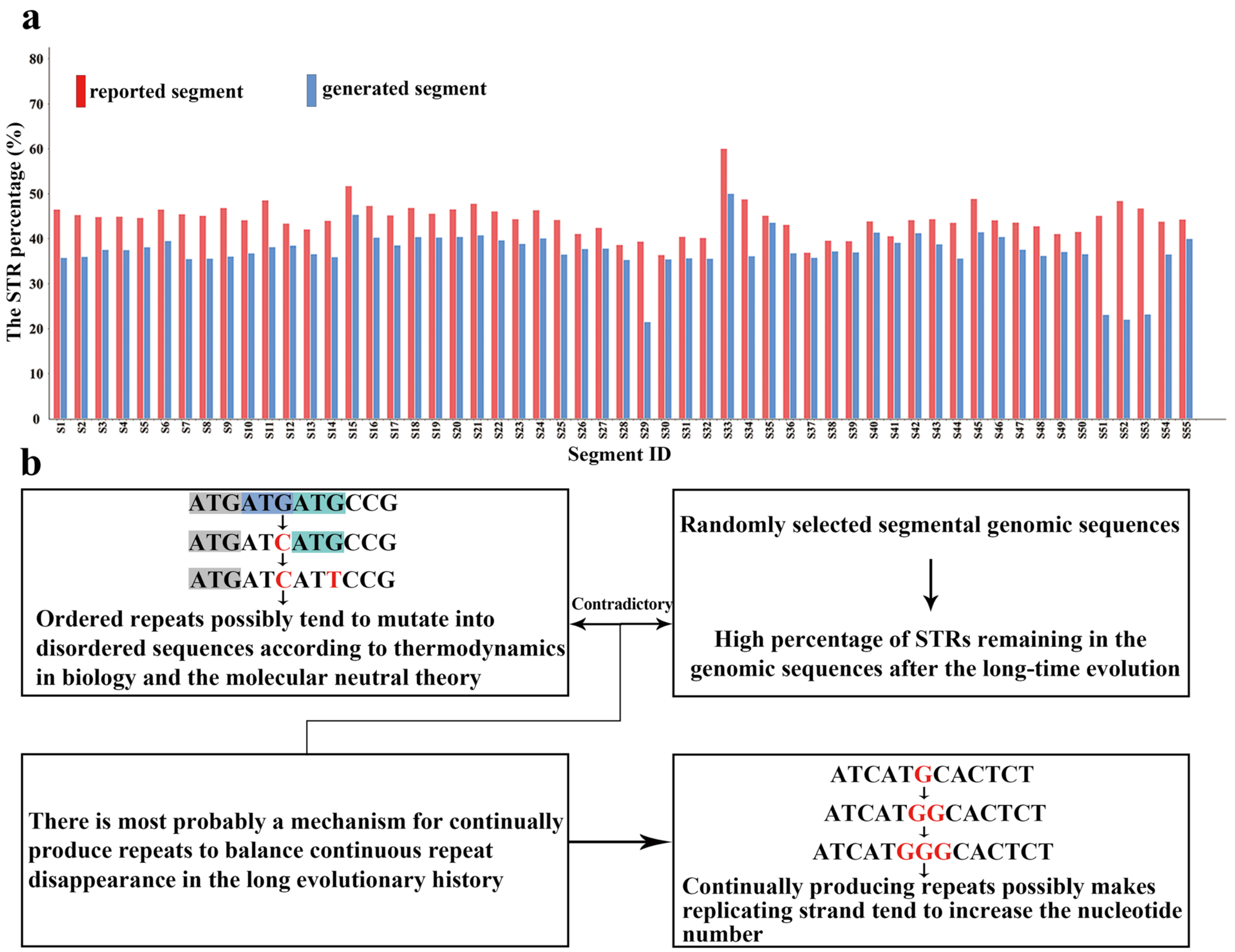


Fig. 1 A high percentage of STRs in genomes and genomes probably tend to produce repeats. **a** STR percentages of 55 randomly-selected reported segments and the control group, which were the sequences generated with the same nucleotide numbers and components as those of the 55 selected reported segments but the random nucleotide orders by a program written in C language. **b** Contradiction analysis of disappearance and high percentage of STRs in the genomes

accepted that the mutation of genomic sequences occurs continually, persistently and permanently. The neutral molecular evolution and molecular clock theories suggest that the nucleotide substitution is constant over the course of evolution; while the thermodynamics in biology states that an isolated system tend to disorder [44–49]. According to the former stated theories, any ordered sequences such as STRs would mutate into disordered sequences in the long evolutionary history without the presence of selective pressure. This theory alone would result in the dilution of STRs and cannot explain the universal presence of preserved high content of STRs in genomes. Therefore, there is most probably an unexplored alternative mechanism for continually producing repeats to balance the continuous disappearance of STRs by random mutation, so as to maintain a high content of short repeat sequences in genomes across all biological kingdoms (Fig. 1b).

Furthermore, the STRs of small iteration numbers were observed to occur more frequently than those of large iteration numbers in all analyzed segments (Table 1, Table S2). A plausible explanation is that the STRs of small iteration numbers may be the basis for forming the STRs of large iteration numbers, otherwise, the STRs of large iteration numbers should occur as frequently as the STRs of small iteration numbers. Some of the longer STRs also possibly mutate into short STRs by contraction and point mutation as debated by many evolutionists [5, 13, 50], and these debates are possible because most short repeats were not considered in their statistics. On the contrary, our observations support the hypothesis that most of longer STRs evolved from the short STRs by expansion, and the genomes tend to produce short repeats by a continual mechanism with the preference of expansion against contraction.

Table 1 The lengths (bp) of STRs with different repeat unit types and different iterations in the segment of the reported human reference X chromosomal sequence at the location of 144,822–231,384 bp

Iteration	Mono[a]	Di	Tri	Tetra	Penta	Hexa	Total
I_{2}	(18128)[b]	10,040	3540	2056	1250	480	17,366
I_{3}	9702	1782	288	156	45	18	11,991
I_{4}	3844	368	12	112	–	–	4336
I_{5}	2095	120	15	20	–	–	2250
I_{6}	600	24	18	0	–	–	642
I_{7}	182	14	-[c]	28	–	–	224
I_{8}	128	16	–	0	–	–	144
I_{9}	54	18	–	36	–	–	108
I_{10}	50	0	–	–	–	–	50
I_{11}	55	22	–	–	–	–	77
I_{12}	24	–	–	–	–	–	24
I_{13}	65	–	–	–	–	–	65
I_{14}	56	–	–	–	–	–	56
I_{15}	45	–	–	–	–	–	45
I_{16}	64	–	–	–	–	–	64
I_{17}	0	–	–	–	–	–	0
I_{18}	36	–	–	–	–	–	36
I_{19}	19	–	–	–	–	–	19
I_{20}	0	–	–	–	–	–	0
I_{21}	42	–	–	–	–	–	42
I_{22}	0	–	–	–	–	–	0
I_{23}	23	–	–	–	–	–	23
I_{24}	0	–	–	–	–	–	0
I_{25}	25	–	–	–	–	–	25
I_{26}	–	–	–	–	–	–	–
I_{27}	–	–	–	–	–	–	–
I_{28}	–	–	–	–	–	–	–
Sum	17,109	12,404	3873	2408	1295	498	37,587

[a] Mononucleotide repeat (Mono), Dinucleotide repeat (Di), Trinucleotide repeat (Tri), Tetranucleotide repeat (Tetra), Pentanucleotide repeat (Penta), Hexanucleotide repeat (Hexa)

[b] The length of mononucleotide repeats with iterations of 2 was not included in this statistics and just used as the reference here

[c] Beyond the largest iteration of this repeat unit type in corresponding analyzed segments were expressed as "-"

Relatively semi-conservative replication

It is well known that each base pair of DNA is a one-to-one correspondence without other extra residue during replication in the double-helix model [38, 39]. And Meselson and Stahl have verified that the replication of DNA chains is semi-conservative by sedimentation techniques based on the diversity differential of DNA with different isotopes, implicating that the number of nucleotides in the replicating strand is consistent with that in the template strand during a complete replication process [51]. However, if the preserved high content of short repeats is produced during replication as described above, the number of nucleotides in the replication strand would be one or several nucleotides/motifs higher than that in the template strand. In vitro experiments also revealed the presence of repeats during DNA replication, and the nascent replication chain has an increase in the number of nucleobases [30, 40, 41, 52]. In the case of our relatively semi-conservative replication model, the replication process can be described as the following formula:

$$N_i = \text{int}[N_0(1+f_1\lambda_1)(1+f_2\lambda_2)\ldots(1+f_i\lambda_i)] \qquad (1)$$

$$\begin{aligned}\Delta N_i &= N_i - N_{i-1}\\ &= \text{int}[N_0 f_i\lambda_i(1+f_1\lambda_1)(1+f_2\lambda_2)\ldots(1+f_{i-1}\lambda_{i-1})] \geq 0\end{aligned} \qquad (2)$$

N_0: The number of nucleotides in the initial template strand;

N_i: The number of nucleotides in the replicating strand during No. i round replication;

int[]: Round the value to the lower integer;

ΔN_i: The difference of the nucleotide numbers between N_i and N_{i-1};

λ_i ($\lambda_i \rightarrow 0$): The coefficient of occurring repeats during No. i round replication; and is most probably an infinitesimal relating to the possibility of repeat occurrence;

f_i ($0 \leq f_i \leq 1$): The fixation coefficient of repeat sequences during No. i round replication.

In general, the number of nucleotides in the replicating strand is likely to have exactly equal to that in the template strand. This observation is consistent with our model when the observed template strand is short and the number of replication rounds is relatively low. For example, the total number of nucleotides in the initial template strand for stable PCR is up to two to three thousand nucleotides. When we suppose $N_0 = 3000$, $\lambda_1 = 10^{-5}$, $f_1 = 1$, the value of ΔN_1 would be 0 according to the formula (2), and therefore, $N_1 = N_0$, causing the replicating strand to be no longer (or no shorter) than the template strand, and the discovery of nascent repeat is unavailable. Nevertheless, when the observed strand is long enough to result in a ΔN_i of larger than 1, our model would explain how the number of nucleotides in the replicating strand changes from that in the template strand. For instance, when we suppose $N_0 = 10^6$, $\lambda_1 = 10^{-5}$, $f_1 = 1$, the value of ΔN_1 would be 10, which could result in the increase of 10 nucleotides (or repeat-motifs) in the replicating strands when compared with the template strand. The increased number of nucleotides may represent nascent repeat sequences according to our relatively semi-conservative replication model.

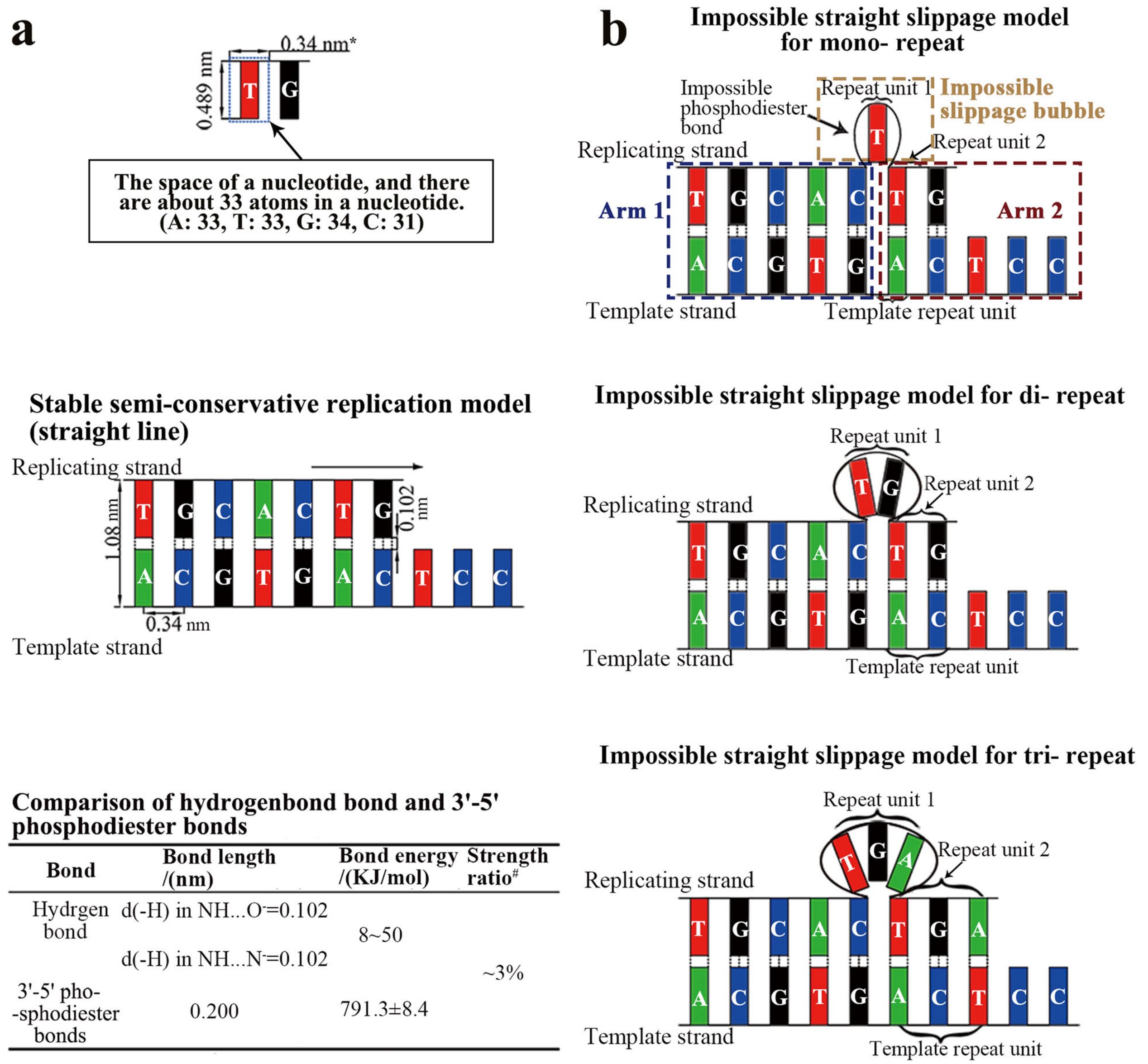
The occurrence of STRs would possibly encounter selective pressure, though it may be different in coding or non-coding regions. We use f_i to represent the fixation possibility of the nascent repeats under selective pressure. A fixation coefficient of 0 ($f_i = 0$) indicates the occurrence of nascent repeats that are lethal mutations and unable to produce survivable offspring, or may be excluded by the DNA repair system [1, 53]. A fixation coefficient of $0 < f_i < 1$ indicates that the nascent STRs are deleterious but still can be fixed in the genome with survived offspring, like Huntington's disease [14]. A fixation coefficient of $0 \le f_i \le 1$ also includes the cases with occurrences of nascent STRs being neutral mutations, which can be either retained or excluded depending on genetic drift. A fixation coefficient of 1 ($f_i = 1$) indicates beneficial mutations, representing that the nascent STRs may help the organism surviving. Therefore, the preserved high content of short repeats suggests that the replicating process frequently produce short repeat sequences which may be fixed neutrally, beneficially, or deleteriously with diseases. This suggests that the replication process may be relatively semi-conservative.

Folded slippage model

The nucleotide chains of various species tend to produce simple repeats, which is likely to be caused by the insertion of additional nucleotides during the replication process. However, the mechanism by which simple repeats actually form during the replication process is still highly debated [5, 50, 54]. The widely accepted mechanism of occurring STR is the replication slippage model, which could explain the expansion and contraction of longer STRs, but not the expansion and contraction of much amounts of short repeats. The existing slippage model is indeed a straight template strand model, with no plausible consideration regarding the space required for the nascent nucleobase, the much stronger phosphodiester bonds when compared with hydrogen bonds (Fig. 2a) [55, 56], and the force that drives the replicate strand slippage. The straight replication slippage model suggests that the STRs possibly occurred by slippage occasionally [13, 58–60], but is rather ambiguous about further details in the mechanism. Actually, there are about 33 atoms in a nucleotide (A: 33, T: 33, G: 34, C: 31) [61], which possess a certain physical space in the molecule. According to previous reports, we simplified a nucleotide space into an intuitive plane model, whose length is about 0.489 nm (length = (distance between the double helix 1.08 - Hydrogen bond length 0.102) / 2), and with a width of 0.34 nm which is the distance between each pair of bases (Fig. 2a) [55–57]. We reconstructed the linear replication slippage model with a CAD geometric calculation by considering the space of bases (Fig. 2b, Fig. S1). If the slippage bubble has enough geometric space to accommodate the repeat unit, the phosphodiester bond would be stretched to far more than 0.34 nm. This is contradictory to the chemical principle that the phosphodiester bonds in DNA is actually much stronger than the hydrogen bonds (Fig. 2a) [57]. Since it is impossible to form a slippage bubble by a larger elongation of the phosphodiester bonds to accommodate the nascent repeat unit, the straight slippage model is insufficient to explain to the occurrence of short repeats and a more sophisticated slippage model should be proposed.

Actually existing replication slippage studies has largely overlooked the validity of the straight template strand assumption in the replication process – the template strands are thought to be perfectly straight in all replication models. Though the template strands are indeed straight in general condition, the possibility of a kinked strand cannot be ruled out. It is well known that the dimension of fully unfolded and extended genomic DNA chains are several magnitudes higher than the dimension of the nucleus (Fig. 3a). For example, the total length of human genome is about 2 m (2×10^9 nm), while the diameter of nucleus is beneath 10^5 nm in human cell [61]. Therefore, the genomic DNA chains are generally highly compacted and folded in the nucleus. During the semi-conservative replication, the replicating molecule is believed to be a straight molecule [40–43], while the replicating enzyme complexes usually straighten the template strand and make the replicating strand well paired with the template strand [40, 62, 63]. However, environmental factors such as temperature, viral proteins or diseases may disrupt the normal works of the enzyme complexes. We speculate that such disruption of the enzyme complex may cause both the replicating strand and the template strand to regain their curved or folded state, resulting in the emergence of provisional kinked strands.

First, we proposed a curved template slippage model for the replication process. When the curved DNA strand is used as the template strand on the inner side, the replication strand is longer than the template strand and can form more nucleotides than the template strand on the outer side. The replication strand should be longer than the template strand so as to provide extra spaces for accommodating the extra repeat bases (Fig. 3b). The links of base pairs mainly depend on 2 types of hydrogen bonds, N—H …: N and N—H …: O [55], with a strength at about 3% of the 3′, 5′-phosphodiester bonds [56, 57, 64, 65] (Fig. 2a). While the distance between the bases is fixed at the backbone, the strengths of the hydrogen bonds are negatively correlated to the distance between every base pair. Therefore, the curved template slippage model would cause the hydrogen bonds to exceed the threshold of 0.167 nm and break off [55]. The curved slippage model partially explains the spaces that form slippage bubble, yet at the cost of forming unstable hydrogen bonds double-chain structures (Arm1 and Arm2) on both sides of the slippage bubble (Fig. 3b, Fig. S2). The curved slippage


a
0.34 nm*
0.489 nm
T
G
The space of a nucleotide, and there are about 33 atoms in a nucleotide. (A: 33, T: 33, G: 34, C: 31)
Stable semi-conservative replication model (straight line)
Replicating strand
1.08 nm
0.102 nm
0.34 nm
Template strand
Comparison of hydrogenbond bond and 3'-5' phosphodiester bonds
b
Impossible straight slippage model for mono- repeat
Impossible phosphodiester bond
Repeat unit 1
Impossible slippage bubble
Repeat unit 2
Replicating strand
Arm 1
Arm 2
Template strand
Template repeat unit
Impossible straight slippage model for di- repeat
Repeat unit 1
Repeat unit 2
Replicating strand
Template strand
Template repeat unit
Impossible straight slippage model for tri- repeat
Repeat unit 1
Repeat unit 2
Replicating strand
Template strand
Template repeat unit

Bond	Bond length /(nm)	Bond energy /(KJ/mol)	Strength ratio#
Hydrgen bond	d(-H) in NH...O^-=0.102 d(-H) in NH...N^-=0.102	8~50	~3%
3'-5' pho--sphodiester bonds	0.200	791.3±8.4	

Fig. 2 Straight strand models of semi-conservative replication and slippage. **a** The space of a nucleotide was drawn. * indicates that those number is the theoretical values (top); The stable straight model of semi-conservative replication (middle); The comparison of hydrogen bond and 3′-5′ phosphodiester bonds (bottom) [55–57]. # indicates the strength ratio was calculated by the strength of hydrogen bond dividing that of phosphodiester bond. **b** The impossible straight slippage models of mononucleotide, dinucleotide and trinucleotide repeats according to the strict geometric calculation of the space of a nucleotide and the stability of hydrogen and phosphodiester bonds

model is an advance from the classic straight slippage model but still has fundamental flaw.

Then we proposed a folded slippage model. The folded template strand forms a slippage bubble above the folding site to accommodate the repeat nucleotides during the replication process. The phosphodiester bonds are fixed and the bases are well paired with stable hydrogen bonds on both sides of the slippage bubble (Fig. 4). With proper folding angle, a stable double-stranded folded slippage structure can provide chances to produce repeats, while satisfying factors including sufficient nucleotide geometric spaces, stable phosphodiester bonds and stable hydrogen bonds. Actually, there are two variations of the folded slippage models: When template strand is on the inner side, the repeat unit duplicates to produce new repetitive unit or repeat expansion (Fig. 4); and when the template strand is on the outer side, the replication strand may make the repetitive sequences to contract (Fig. 5). The features of this folded slippage model can explain the widely observed STR mutations with expansion and contraction of repeat

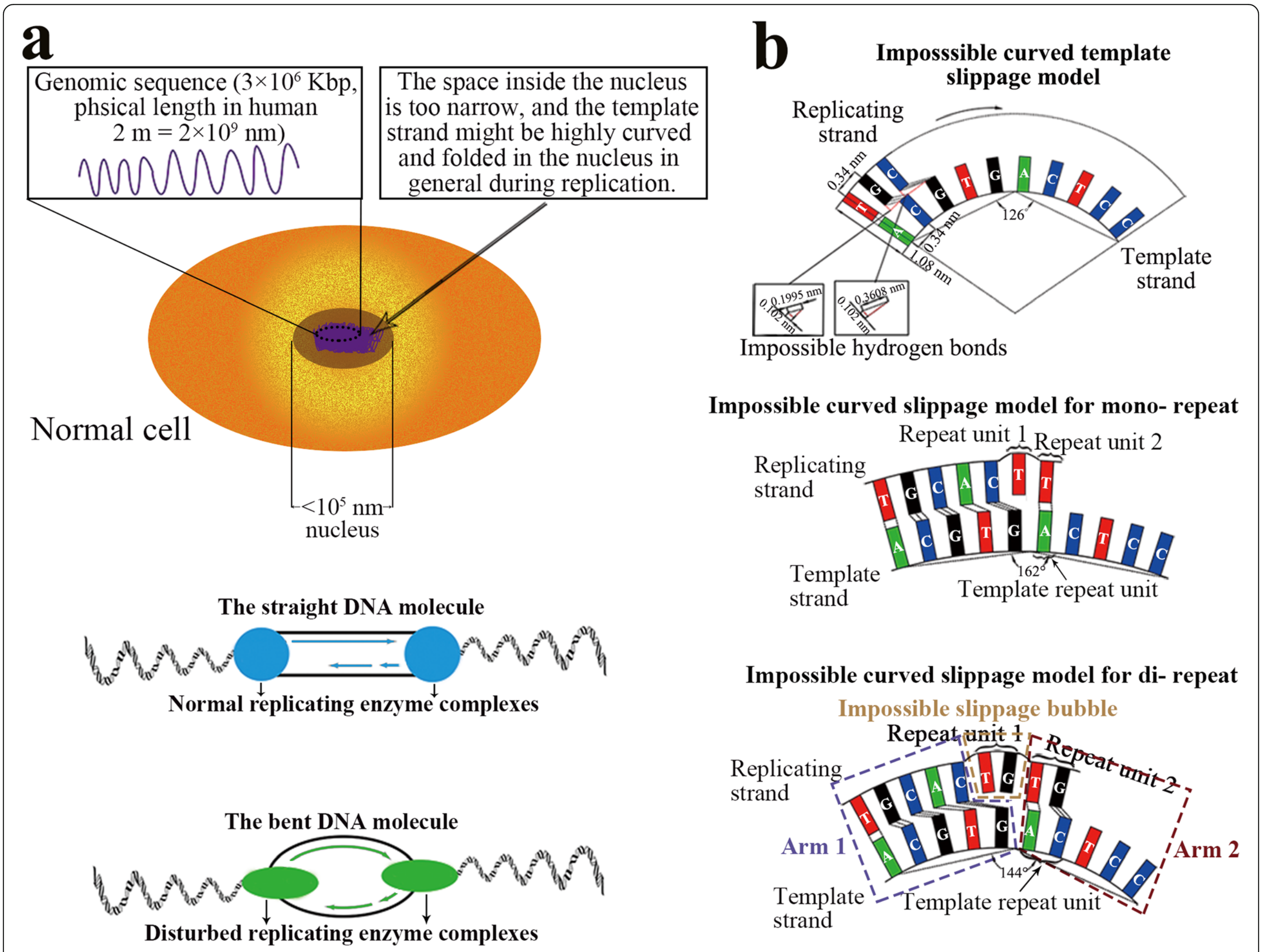

a
Genomic sequence (3×10⁶ Kbp, phsical length in human 2 m = 2×10⁹ nm)
The space inside the nucleus is too narrow, and the template strand might be highly curved and folded in the nucleus in general during replication.
Normal cell
<10⁵ nm nucleus
The straight DNA molecule
Normal replicating enzyme complexes
The bent DNA molecule
Disturbed replicating enzyme complexes
b
Imposssible curved template slippage model
Replicating strand
0.34 nm
T G C
A C G T G A C T C C
0.34 nm
1.08 nm
126°
Template strand
0.1995 nm
0.102 nm
0.3608 nm
0.102 nm
Impossible hydrogen bonds
Impossible curved slippage model for mono- repeat
Repeat unit 1
Repeat unit 2
Replicating strand
T G C A C T T
A C G T G A C T C C
Template strand
162°
Template repeat unit
Impossible curved slippage model for di- repeat
Impossible slippage bubble
Repeat unit 1
Repeat unit 2
Replicating strand
T G C A C T G T G
A C G T G A C T C C
Arm 1
Arm 2
144°
Template strand
Template repeat unit

Fig. 3 The DNA chain is highly curved or folded in the nucleus and the impossible curved slippage model. **a** Schematic diagram of the size of the nuclear space (top) [61]; The normal replicating enzymes complex straighten the DNA chain, while the disturbed replicating enzymes complex may cause the DNA molecule return to curved state (bottom). **b** Impossible curved template slippage model according to the strict geometric calculation of the space of a nucleotide and the stability of hydrogen and phosphodiester bonds (top); Mono- and dinucleotide repeats may be impossibly produced in curved replicating strands (middle and bottom)

units [5, 13, 50, 59, 66]. In addition, replication slippage of template strands with different folding angles may result in the expansion or contraction of repeat units with different sizes. When template chains are folded on the inner side at a folding angle of 18°, 36°, 54°, 72°, 90° and 108°, the replication strands would produce mono-, di-, tri-, tetra-, penta-, hexa-nucleotide repeat expansions, respectively (Fig. 4). With fixed phosphodiester bond, it is necessary to break off more hydrogen bonds to produce higher number of repeats. For example, since 2 or 3 hydrogen bonds are used to stabilize each base pair, 12–18 hydrogen bonds need to be broken to produce hexanucleotide repeats. This suggested that the difficulty to form repeats from mono- to hexanucleotide gradually increases, which explains our statistic data in which the occurrence of mono-, di-, tri-, tetra-, penta- and hexanucleotide repeat gradually decreases (Table 1, Table S2). Similarly, when template chains are folded on the outer side at a rotation angel of 18°, 36°, 54°, 72°, 90° and 108°, the replication strands will produce corresponding repeat contractions respectively (Fig. 5). These features of our folded slippage model can explain the emergence of short tandem repeats which usually refers to the tandem repeats with repeat units from mono- to hexanucleotides [5, 22, 27]. According to this rule, we also describe the possible folded template slippage models of hepta-, octa-, nona- and deca-nucleotide repeats (Figs. S3 and S4), while the replicating strand must break off 14–21, 16–24, 18–27, 20–30 hydrogen bonds to make a folded slippage bubble, respectively. Such long tandem repetitive sequences are unlikely to occur since the energy to break off 14–30 hydrogen bonds are on the same scale as the energy to break off one phosphodiester bond, which explains the observations that they are often much less abundant in the genomes [59, 67]. Our folded slippage model can also explain how the (A_mT_n) repeats tend to grow faster than (G_mC_n) repeats

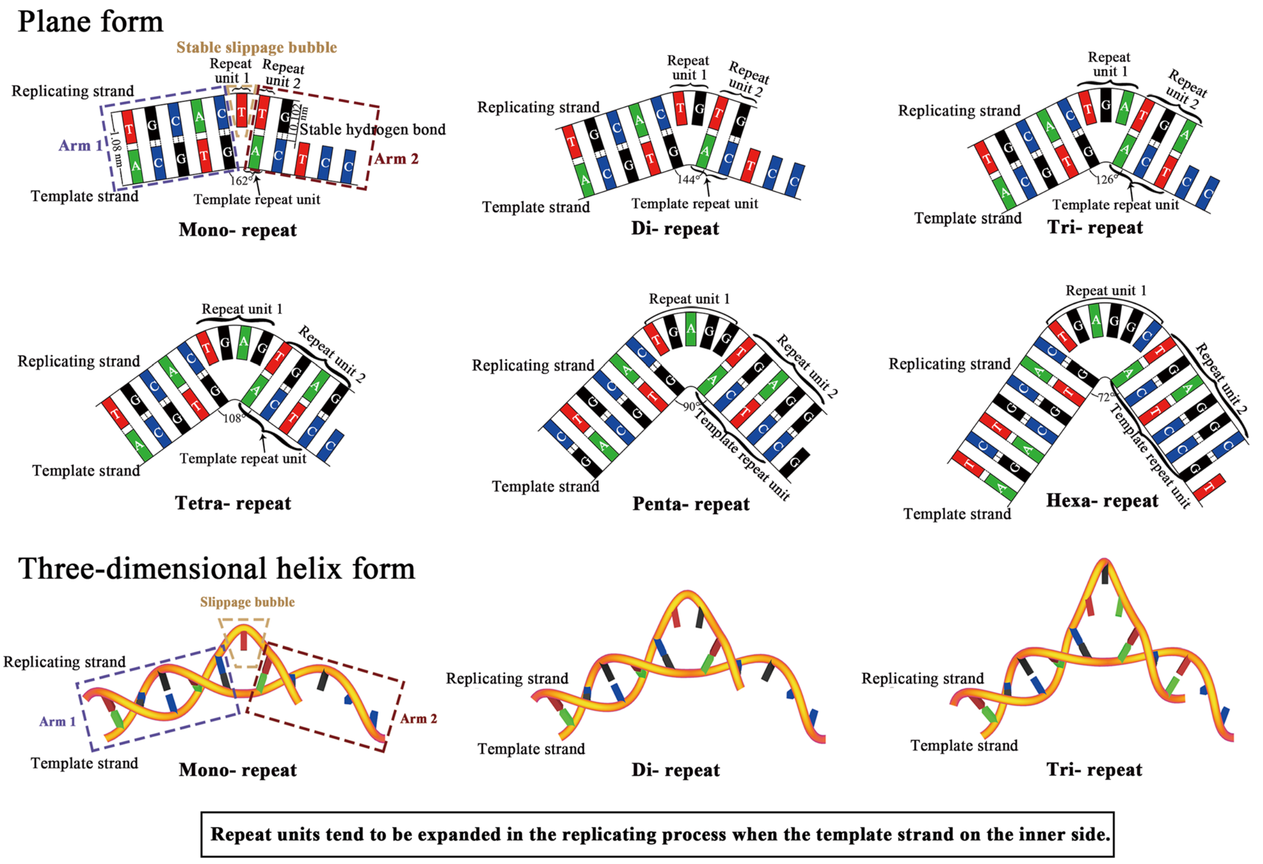


Fig. 4 Stable folded slippage models of mononucleotide to hexanucleotide repeats amplification according to the strict geometric calculation of the space of a nucleotide and the stability of hydrogen and phosphodiester bonds. Repeat units tend to be expanded in the replicating strands when the template strands are on the inner side of the folded slippage models respectively. The bottom 3 sub-figures were the folded slippage models in three-dimensional helix form

because smaller number of broken hydrogen bonds in the (A_mT_n) repeats impose lower energy barrier for repeat expansion [21, 37, 68, 69]. Although this folded slippage model is a simplified model described in a plane form, it simulates and explains the repeat sequences producing process. We also build a simplified double-helical model in three-dimensional forms to show the folded slippage model more intuitively (Figs. 4 and 5), while the precise folding angle and other issues deserve further study.

When compared with the straight template slippage model, the folded template model exhibits enough geometric space in the slippage bubble to accommodate repeat nucleotides without stretching the phosphodiester bonds. When compared with the curved template model, the folded model has two sides of the slippage bubble stably paired, and has Arm1 and Arm2 similar to the straight template replication model at both sides (Figs. 4 and 5). The folded model takes full account of the space required by nucleotides, the stability of phosphodiester bonds, and the strength comparison between phosphodiester bonds and hydrogen bond. This model can explain STR mutations with repeat unit expansion and contraction, and provides a plausible explanation for the production of short repeats production in the replicating process which otherwise neither the straight slippage model nor the curved slippage model can explain. The folded template strand slippage model may be responsible for the continual production of repeat sequences and the retention of high percentage of repeat sequences in genomes.

Discussion

According to the folded slippage model, the template chain folding on the inner side may make the replicating chain slippage for repeat expansion, while the template chain folding on the outer side may make the replicating chain slippage for repeat contraction. At a first glance, the possibility of repeat expansion and contraction may appear to be the same. However, there are two manners for the repeat sequences contraction, one is above mentioned the template chain folds on outside,

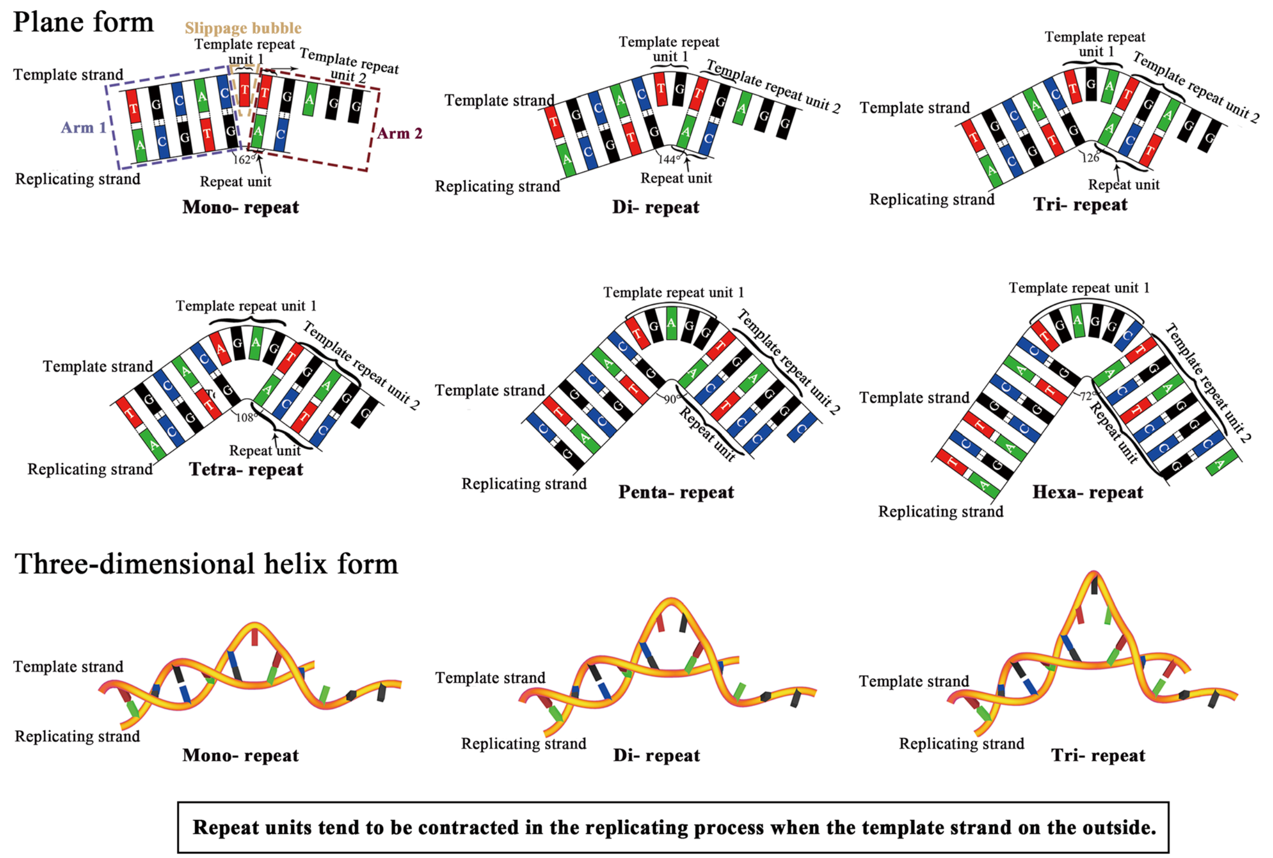


Fig. 5 Stable folded slippage models of mononucleotide to hexanucleotide repeats contraction according to the strict geometric calculation of the space of a nucleotide and the stability of hydrogen and phosphodiester bonds. Repeat units tend to be subtracted in the replicating strands when the template strands are on the outside of the folded slippage models respectively. The bottom 3 sub-figures were the folded slippage models in three-dimensional helix form

another is also general mutations stated above. The high content of repeat sequences is still in a stable state in the genome of each species, implicating a higher rate for repeat expansion when compared with repeat contraction, which is also reported in many other studies [30, 52, 70].

Under normal circumstances, the replicating enzyme complexes may provide power for balancing the external forces to drag the template DNA molecule straight. However, when the replicating enzyme complexes are disturbed, the replicating straight template DNA chain should return to folded under external forces from the narrow and crowded cell nucleus. We proposed an external force model for template strand returning to folded, and this model may be helpful to explore the probability of expansion and contraction. When the template strand is on the inner side, the nucleotide bases point outward, and the space of bases at the folded site become wide and loose at outward part; while it is on the outer side, the base in the folding position is squeezed inward. Comprehensive consideration of the small difference of the space of nucleotides at the folded site reveals that the external forces to make template strand folded with bases loose should be smaller than that with base squeezed. Therefore, the external force required for the template strand folded on the outside (F^o) is inevitable greater than that on the inner side (F^i). $F^o > F^i$ suggests that the probability for the template strand folded on the inner side is higher than that on the outer side. Our folded slippage model suggested that the repeats tend to expand when the template strand is on inner side and tend to contract when the template strand is on the outer side. Therefore, the odds of repeat expansion (P^e) is higher than that for repeat contraction (P^c), which can be described as $P^e > P^c$ (Fig. 6). The STR studies, like in Huntington disease related locus and myotonic dystrophy type 1 locus, all showed STR expansion biased [13, 14, 71–73], which proves that the expansion of short STRs are more frequent than that of contraction.

Thus, according to formula (2):

When the template strand on the outer side, repeats tend to contract, so $\lambda^c < 0$,

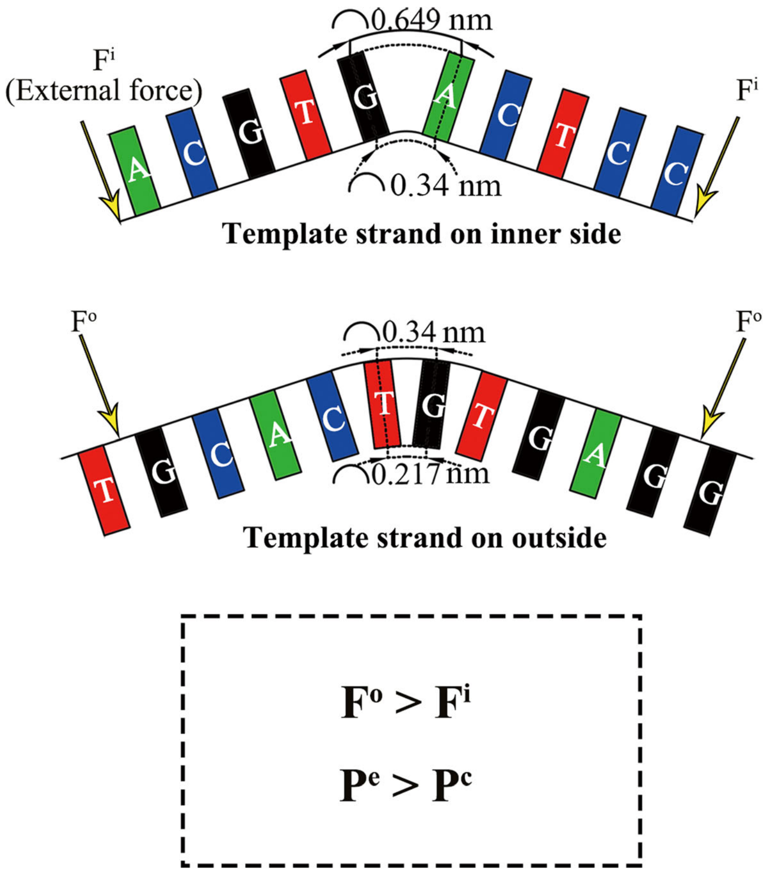


Fig. 6 Repeat production incline to expansion. F^O, F^i refer to the force required for the two template strands to bend, respectively. $F^O > F^i$ means that the force of the template strand bending downward is greater than the bending upward, and $P^e > P^C$ means that the possibility of the template strand bending upward is greater than the downward bending

thus, $\Delta N^c = N^c{}_i - N^c{}_{i-1} = \mathrm{int}[N_0 f^c{}_i \lambda^c{}_i (1 + f^c{}_1 \lambda^c{}_1)(1 + f^c{}_2 \lambda^c{}_2)\ldots(1 + f^c{}_{i-1} \lambda^c{}_{i-1})] \le 0$.

When the template strand on the inner side, repeats tend to expand, so $\lambda^e > 0$,

thus, $\Delta N^e = N^e{}_j - N^e{}_{j-1} = \mathrm{int}[N_0 f^e{}_j \lambda^e{}_j (1 + f^e{}_1 \lambda^e{}_1)(1 + f^e{}_2 \lambda^e{}_2)\ldots(1 + f^e{}_{j-1} \lambda^e{}_{j-1})] \ge 0$.

The general repeat expansion and contraction can be described as:

Because λ was defined as coefficient of occurring repeats, the possibility of repeat expansion (P^e) is positively proportional to λ^e and the possibility of contraction (P^c) is positively proportional to the absolute value of λ^c ($|\lambda^c|$). Under the assumptions that $f^e = f^c = f$, $i = j$, and as generally $P^e > P^c$, then $\lambda^e > |\lambda^c|$, and also $\Sigma[|\lambda^e{}_j(1 + f\lambda^e{}_1)(1 + f\lambda^e{}_2)\ldots(1 + f\lambda^e{}_{j-1})|] \ge \Sigma\,[|\lambda^c{}_i(1 + f\lambda^c{}_1)(1 + f\lambda^c{}_2)\ldots(1 + f\lambda^c{}_{i-1})|]$,

therefore, $\Sigma\Delta N = |\Sigma\Delta N^e| - |\Sigma\Delta N^c| \ge 0$.

So, when the external forces for returning the folded template strand were considered, the possibility of repeat expansion should be higher than that of repeat contraction, then the revised formula (2) is also able to explain the retention of high percentage of short repeats in genomes under a mechanism of continually producing repeats. This mechanism might result from the folded template chain slippage model, which is possibly responsible for the widely occurring STRs in eukaryotic,

$$\left|\sum \Delta N^e\right| = \left|\ \mathrm{int}\left[\sum N_0 f^e{}_j \lambda^e{}_j (1 + f^e{}_1 \lambda^e{}_1)(1 + f^e{}_2 \lambda^e{}_2)\ldots\left(1 + f^e{}_{j-1} \lambda^e{}_{j-1}\right)\right]\right|;$$

$$\left|\sum \Delta N^c\right| = \left|\ \mathrm{int}\left[\sum N_0 f^c{}_i \lambda^c{}_i (1 + f^c{}_1 \lambda^c{}_1)(1 + f^c{}_2 \lambda^c{}_2)\ldots(1 + f^c{}_{i-1} \lambda^c{}_{i-1})\right]\right|;$$

$$\sum \Delta N = \left|\sum \Delta N^e\right| - \left|\sum \Delta N^c\right| = \ \mathrm{int}\left[N_0 \sum\left[\left|f^e{}_j \lambda^e{}_j (1 + f^e{}_1 \lambda^e{}_1)(1 + f^e{}_2 \lambda^e{}_2)\ldots\left(1 + f^e{}_{j-1} \lambda^e{}_{j-1}\right)\right| - |f^c{}_i \lambda^c{}_i (1 + f^c{}_1 \lambda^c{}_1)(1 + f^c{}_2 \lambda^c{}_2)\ldots(1 + f\lambda^c{}_{i-1} \lambda^c{}_{i-1})|\right]\right].$$

prokaryotic and also viral genomes. We improved the straight slippage model to a folded slippage model by fully considering the geometric spaces of nucleotide bases, the relationship between phosphodiester and hydrogen bond, and the stability of these bonds. The slippage model showed that the straight replicating template DNA may partially regain its folded state resulting from disturbed replicating enzyme complexes, and may provide chances for continually producing much amount of short repeats; though the long unit repeats may be explained by the former slippage model [33, 59].

The easily forming of folded slippage may also be responsible for the widely observed fact that repetitive part of genome is usually evolved one hundred or more times than other parts with only repeat unit expansion and contraction [1, 18, 50, 74], though the repeats occurred more in non-coding regions than in coding regions possibly due to different selective pressures [5, 13, 59]. Most of the emerging repeats should be lethal mutation and may have been negatively selected to lost; some of emerging repeats should be deleterious in genomes and responsible for a series of diseases [72, 73, 75, 76]; many neutral repeat expansions may be lost or fixed with no functions in genomes by genetic drift [77]; and some beneficial repeat expansions may promote the emergence of different new properties or functions – all of which lead to the abundance of repeat sequences in the genomes with a diversified set of roles as reported in the literature [9–11, 66, 68, 78, 79]. The longer repeats might originate from continuous short repeat expansion by the folded template slippage; the longer genomes possibly evolved from the short genomes in the long evolutionary replicating process.

Conclusions

The universal presence of high-content short repeats is possibly a common characteristic of genomes across all biological kingdoms, which indicates a mechanism for continuous production of repeats. We proposed a folded replication slippage model, which provides a reasonable explanation for the continuous occurrences of STRs and their high contents in genomes with improving the existing straight-line slippage model, and this folded replication slippage model also suggests that expansion exists more commonly than contraction in the STRs without the presence of selective pressure. This model also contributes to the explanation of STR-to-genome evolution and is an alternative model that complements semi-conservative replication.

Methods

Sequences resource

We randomly selected 50 species covering animals, plants, fungus, protozoa, bacteria, archaea and viruses, according to the list of "KEGG Organism: Complete Genome" [80]. To simplify the analyses and make the analyzed data statistically representative, we randomly chose 55 sequence segments with size range from 3000 to 96,600 bp; the segments are out of 55 full genomic sequences from the 50 selected species, in which 5 species were randomly selected with double genomic sequences and 45 species were randomly selected with single genomic sequence from the reported data in Genbank; the segments were selected randomly in position and avoided to select incompletely sequenced gaps; the accession numbers with the related information were listed in Table S1.

Repeat extraction

The perfect simple sequence repeats were extracted by Imperfect Microsatellite Extraction Webserver [81] from those 55 randomly selected segments. The minimum iterations for all perfect mono- to hexanucleotide repeats were set at 3, 2, 2, 2, 2, 2 to mine the data more completely in this study, comparing with most researchers setting iterations at relatively higher self-defined values, and 3 iterations for mononucleotide repeats were defined to ensure to be commonly recognized as the STRs.

Null hypothesis test

We also extracted perfect mono- to hexanucleotide repeats under the above threshold in the sequences that were generated by a program written in C language (Program S1). The nucleotide compositions and numbers of the generate segments were the same as those of the selected segments, however, the nucleotide orders of the generate segments were randomly rearranged in the C program. Then, the validating test, which can verify that the short STRs extracted in those 55 reported segments are not randomly occurred, was based on the comparison of the STR percentages in the reported segments and the generated segments.

Model drawing of DNA replication

Different models were drawn to simulate the DNA replication. Normally in straight model, the hydrogen bond length between 2 paired nucleotides is reported to be 0.102 nm and the distance between 2 neighboring nucleotides is 0.34 nm, importantly, owing to the nucleotides occupying almost same space in DNA strands, the space of a nucleotide was simplified into a geometric plane form in this analysis, which was 0.489 nm in length and 0.34 nm in width. Then we applied AutoCAD [82] to draw the straight, curved and folded slippage models according to the strict geometric calculation of the spaces of nucleotides and different strengths between hydrogen bonds and phosphodiester bonds. And the slippage models in helix structure were achieved by Rhino [83], which is an industrial drawing software.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-06949-5.

Additional file 1: Table S1. The basic segment information and percentage comparison of short repeats between the randomly selected segments and the comparison of short repeats between the randomly selected segments and the corresponding generated segments.

Additional file 2: Table S2. The total lengths (bp) of STR with different repeat units and different iterations in 55 analyzed segmental sequences under the standard of 3, 2, 2, 2, 2, 2

Additional file 3. Supplementary Program

Additional file 4.

Additional file 5.

Additional file 6.

Additional file 7.

Abbreviations

SSR: Simple sequence repeat; SSRs: Simple sequence repeats; STR: Short tandem repeat; STRs: Short tandem repeats

Acknowledgments

The authors thank editor and anonymous reviewers for suggestions on the manuscript. The authors thank Heping Zheng for language editing, who is previous researcher (2001-2019) in University of Virginia, USA. The authors thank Qijun Tian for his technological help at designing the folded slippage model in helix structure, who is from School of design, Hunan University, China.

Authors' contributions

ZT designed and directed this study. HZ, DL, and XZ performed the data analysis and model drawing. XW, SP, HH and RS collected sequence data and discussed model drawing. ZT, HZ and DL prepared the manuscript. The author(s) read and approved the final manuscript.

Funding

This work was jointly supported by funding from the "National Key Plan for Scientific Research and Development of China" (grant No. 2016YFC1200200 and 2016YFD0500300).

Availability of data and materials

The information including accession number of 55 analyzed sequence segments is listed in Table S1 and also available in the following git-hub web link. Table S2 is the dataset of STRs with different repeat units and different iterations in 55 analyzed segmental sequences under the standard of 3, 2, 2, 2, 2, 2, which are also available in the following git-hub web link. Program S1 for generating segments, Table S1 and Table S2 can be downloaded from https://github.com/DooYal/Supplementary-materials-for-submitting-relatively-...-.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 13 April 2020 Accepted: 27 July 2020
Published online: 17 August 2020

References

1. Mandal R, Samstein RM, Lee KW, Havel JJ, Wang H, Krishna C, Sabio EY, Makarov V, Kuo FS, Blecua P, et al. CANCER genetic diversity of tumors with mismatch repair deficiency influences anti-PD-1 immunotherapy response. Science. 2019;364(6439):485–91.
2. Zhao X, Tian Y, Yang R, Feng H, Ouyang Q, Tian Y, Tan Z, Li M, Niu Y, Jiang J. Coevolution between simple sequence repeats (SSRs) and virus genome size. BMC Genomics. 2012;13(1):435.
3. Chen M, Tan Z, Zeng G, Peng J. Comprehensive analysis of simple sequence repeats in pre-miRNAs. Mol Biol Evol. 2010;27(10):2227–32.
4. Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. Unstable tandem repeats in promoters confer transcriptional evolvability. Science. 2009;324(5931):1213–36.
5. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004;5(6):435–45.
6. Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet. 2002;30(2): 194–200.
7. Ince AG, Karaca M, Onus AN. CAPS-microsatellites: use of CAPS method to convert non-polymorphic microsatellites into useful markers. Mol Breed. 2010;25(3):491–9.
8. Lin WH, Kussell E. Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. Nucleic Acids Res. 2012;40(6):2399–413.
9. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol. 2004;21(6):991–1007.
10. Hood DW, Deadman ME, Jennings MP, Bisercic M, Fleischmann RD, Venter JC, Moxon ER. DNA repeats identify novel virulence genes in Haemophilus influenzae. Proc Natl Acad Sci U S A. 1996;93(20):11121–5.
11. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018;19(5):286–98.
12. Jain A, Vale RD. RNA phase transitions in repeat expansion disorders. Nature. 2017;546(7657):243–7.
13. Mirkin SM. Expandable DNA repeats and human disease. Nature. 2007; 447(7147):932–40.
14. Macdonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, Barnes G, Taylor SA, James M, Groot N. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72(6):971–83.
15. Chan EM, Shibue T, McFarland JM, Gaeta B, Ghandi M, Dumont N, Gonzalez A, McPartlan JS, Li TX, Zhang YX, et al. WRN helicase is a synthetic lethal target in microsatellite unstable cancers. Nature. 2019;568(7753):551–6.
16. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. Comprehensive characterization of Cancer driver genes and mutations. Cell. 2018;173(2): 371–85. e318.
17. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014;15(9):585–98.
18. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial Cancer genomes. Cell. 2013;155(4):858–68.
19. Karaca M, Bilgen M, Onus AN, Ince AG, Elmasulu SY. Exact tandem repeats analyzer (E-TRA): a new program for DNA sequence mining. J Genet. 2005; 84(1):49–54.
20. Bilgen M, Karaca M, Onus AN, Ince AG. A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences. Bioinformatics. 2004;20(18):3379–86.
21. Tian XJ, Strassmann JE, Queller DC. Genome nucleotide composition shapes variation in simple sequence repeats. Mol Biol Evol. 2011;28(2):899–909.
22. Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, Makova KD. What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at a/T and GT/AC repeats. Genome Biol Evol. 2010;2:620–35.
23. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
24. Chen M, Tan Z, Zeng G. Microsatellite is an important component of complete hepatitis C virus genomes. Infect Genet Evol. 2011;11(7):1646–54.
25. Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K, Sundaram RM. Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. Bioinformatics. 2007;23(1):1–4.

26. George B, Alam CM, Jain SK, Sharfuddin C, Chakraborty S. Differential distribution and occurrence of simple sequence repeats in diverse geminivirus genomes. Virus Genes. 2012;45(3):556–66.
27. Zhao X, Tan Z, Feng H, Yang R, Li M, Jiang J, Shen G, Yu R. Microsatellites in different Potyvirus genomes: survey and analysis. Gene. 2011;488(1–2):52–6.
28. Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, Lim WK, Ong CK, Chan K, Cheng VKY, et al. The draft genome of tropical fruit durian (Durio zibethinus). Nat Genet. 2017;49(11):1633–41.
29. Hunt VL, Tsai IJ, Coghlan A, Reid AJ, Holroyd N, Foth BJ, Tracey A, Cotton JA, Stanley EJ, Beasley H, et al. The genomic basis of parasitism in the Strongyloides clade of nematodes. Nat Genet. 2016;48(3):299–307.
30. Fungtammasan A, Ananda G, Hile SE, Su MSW, Sun C, Harris R, Medvedev P, Eckert K, Makova KD. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 2015; 25(5):736–49.
31. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014;46(7):707–13.
32. Gadgil R, Barthelemy J, Lewis T, Leffak M. Replication stalling and DNA microsatellite instability. Biophys Chem. 2016;225:38–48.
33. Viguera E, Canceill D, Ehrlich SD. Replication slippage involves DNA polymerase pausing and dissociation. EMBO J. 2001;20(10):2587–95.
34. Huang TY, Chang CK, Kao YF, Chin CH, Ni CW, Hsu HY, Hu NJ, Hsieh LC, Chou SH, Lee IR. Parity-dependent hairpin configurations of repetitive DNA sequence promote slippage associated with DNA expansion. Proc Natl Acad Sci U S A. 2017;114(36):9535–40.
35. Garcia-Diaz M, Bebenek K, Krahn JM, Pedersen LC, Kunkel TA. Structural analysis of strand misalignment during DNA synthesis by a human DNA polymerase. Cell. 2006;124(2):331–42.
36. Lai YL, Sun FZ. The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol. 2003;20(12):2123–31.
37. Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res. 1992;20(2):211–5.
38. Watson JD, Crick FHC. Molecular structure of deoxypentose nucleic acids. Nature. 1953;171:738–40.
39. Watson JD, Crick FHC. Genetical implications of the structure of dexoyribonucleic acid. Nature. 1953;171:964–7.
40. Kiefer JR, Mao C, Braman JC, Beese LS. Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. Nature. 1998;391(6664): 304.
41. Doublié S, Tabor S, Long AM, Richardson CC, Ellenberger T. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 a resolution. Nature. 1998;391(6664):251–8.
42. Bell SD. DNA replication: archaeal oriGINS. BMC Biol. 2011;9:36.
43. Costa A, Ilves I, Tamberg N, Petojevic T, Nogales E, Botchan MR, Berger JM. The structural basis for MCM2-7 helicase activation by GINS and Cdc45. Nat Struct Mol Biol. 2011;18(4):471–7.
44. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267(5608):275–6.
45. Kimura M: **The neutral theory of molecular evolution**. *Sci Am* 1979, **241**(5): 98–100, 102, 108 passim.
46. Margoliash E. Primary structure and evolution of cytochrome C. Proc Natl Acad Sci U S A. 1963;50:672–9.
47. Zuckerkandl E, Pauling LB: **Molecular disease, evolution, and genic heterogeneity**. In: *Horizons in Biochemistry.* Edited by Pullman B, Kasha M, SzentGyörgyi A. New York: Academic Press, New York; 1962: 189–225.
48. Zuckerkandl E, Pauling LB: **Evolutionary divergence and convergence in proteins**. In: *Evolving Genes and Proteins.* Edited by Bryson V, Vogel HJ. New York: Academic Press, New York; 1965: 97–166.
49. Bharadwaj S, Montazeri R, Haynie DT. Direct determination of the thermodynamics of polyelectrolyte complexation and implications thereof for electrostatic layer-by-layer assembly of multilayer films. Langmuir. 2006; 22(14):6093–101.
50. Kelkar YD, Eckert KA, Chiaromonte F, Makova KD. A matter of life or death: how microsatellites emerge in and vanish from the human genome. Genome Res. 2011;21(12):2038–48.
51. Meselson M, Stahl FW. The replication of DNA in Escherichia Coli. Proc Natl Acad Sci U S A. 1958;44(7):671–82.
52. Fungtammasan A, Tomaszkiewicz M, Campos-Sanchez R, Eckert KA, DeGiorgio M, Makova KD. Reverse transcription errors and RNA-DNA differences at short tandem repeats. Mol Biol Evol. 2016;33(10):2744–58.
53. Jeggo PA, Pearl LH, Carr AM. DNA repair, genome stability and cancer: a historical perspective. Nat Rev Cancer. 2015;16(1):35.
54. Torresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019;47(21):10994–1006.
55. Heyrovska R. Dependence of the length of the hydrogen bond on the covalent and cationic radii of hydrogen, and additivity of bonding distances. Chem Phys Lett. 2006;432(1–3):348–51.
56. Gao F, Yin C, Yang P. Coordination chemistrymimics of nuclease-activity in the hydrolytic cleavage of phosphodiester bond. Chin Sci Bull. 2004;49(16): 1667–80.
57. Wang Q. Hydrogen bond in organic chemistry. Tianjin, China: Tianjin University Press; 1993.
58. Leclercq S, Rivals E, Jarne P. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. Genome Biol Evol. 2010;2(4):325–35.
59. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet. 2010;44(1):445–77.
60. Ohshima K, Wells RD. Hairpin formation during DNA synthesis primer realignment in vitro in triplet repeat sequences from human hereditary disease genes. J Biol Chem. 1997;272(27):16798–1806.
61. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular biology of the cell. 4th ed. New York: Garland Science; 2002.
62. Costantino L, Sotiriou SK, Rantala JK, Magin S, Mladenov E, Helleday T, Haber JE, Iliakis G, Kallioniemi OP, Halazonetis TD. Break-induced replication repair of damaged forks induces genomic duplications in human cells. Science. 2014;343(6166):88–91.
63. Fragkos M, Ganier O, Coulombe P, Mechali M. DNA replication origin activation in space and time. Nature Reviews: Molocular Cell Biology. 2015;16(6):360–74.
64. Luo YR. Comprehensive handbook of chemical bond energies. Boca Raton, FL: CRC Press; 2007.
65. Griffiths AJ, Miller JH, Suzuki DT, Lewontin RC. An introduction to genetic analysis. 7th ed; 2000.
66. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016;48(1):22–9.
67. Legendre M, Pochet N, Pak T, Verstrepen KJ. Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res. 2007;17(12): 1787–96.
68. Sinai MIT, Salamon A, Stanleigh N, Goldberg T, Weiss A, Wang YH, Kerem B. AT-dinucleotide rich sequences drive fragile site formation. Nucleic Acids Res. 2019;47(18):9685–95.
69. Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol. 2001;18(7):1161–7.
70. Neil AJ, Liang MU, Khristich AN, Shah KA, Mirkin SM. RNA-DNA hybrids promote the expansion of Friedreich's ataxia (GAA)(n) repeats via break-induced replication. Nucleic Acids Res. 2018;46(7):3487–97.
71. Higham CF, Morales F, Cobbold CA, Haydon DT, Monckton DG. High levels of somatic DNA diversity at the myotonic dystrophy type 1 locus are driven by ultra-frequent expansion and contraction mutations. Hum Mol Genet. 2012;21(11):2450–63.
72. Sznajder LJ, Swanson MS: **Short Tandem Repeat Expansions and RNA-Mediated Pathogenesis in Myotonic Dystrophy**. Int J Mol Sci 2019, **20**(13).
73. Larson E, Fyfe I, Morton AJ, Monckton DG. Age-, tissue- and length-dependent bidirectional somatic CAG*CTG repeat instability in an allelic series of R6/2 Huntington disease mice. Neurobiol Dis. 2015;76:98–111.
74. Giesselmann P, Brandl B, Raimondeau E, Bowen R, Rohrandt C, Tandon R, Kretzmer H, Assum G, Galonska C, Siebert R, et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. Nat Biotechnol. 2019;37(12):1478–81.
75. Arturo LC, Cleary JD, Pearson CE. Repeat instability as the basis for human diseases and as a potential target for therapy. Nat Rev Mol Cell Biol. 2010;11(3):165–70.
76. Sun JH, Zhou LD, Emerson DJ, Phyo SA, Titus KR, Gong WF, Gilgenast TG, Beagan JA, Davidson BL, Tassone F, et al. Disease-associated short tandem repeats co-localize with chromatin domain boundaries. Cell. 2018;175(1): 224–38.
77. Muller MJ, Neugeboren BI, Nelson DR, Murray AW. Genetic drift opposes mutualism during spatial population expansion. Proc Natl Acad Sci U S A. 2014;111(3):1037–42.

78. Mrazek J. Analysis of distribution indicates diverse functions of simple sequence repeats in mycoplasma genomes. Mol Biol Evol. 2006;23(7): 1370–85.
79. Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. Unstable tandem repeats in promoters confer transcriptional evolvability. Science. 2009;324(5931):1213–6.
80. KEGG Organism: Complete Genome. https://www.kegg.jp/kegg/catalog/org_list.html. .
81. Mudunuri SB, Nagarajaram HA. IMEx: Imperfect microsatellite extractor. Bioinformatics. 2007;23(10):1181–7.
82. AutoCAD for Mac & Windows | 2D/3D CAD Software | Autodesk. https://www.autodesk.com.sg/products/autocad/overview. Accessed 15 Dec 2015.
83. Rhino 6 for Windows download. https://www.rhino3d.com/download/rhino-for-windows/6/evaluation. Accessed Mar 1 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.