Relatively semi-conservative replication and a folded slippage model for simple repeat sequences

Hongxi Zhang[a,1], Douyue Li[a,1], Xiangyan Zhao[a,1], Saichao Pan[a,1], Xiaolong Wu[a], Shan Peng[a], Hanrou Huang[a], Ruixue Shi[a] and Zhongyang Tan[a,*]

[a]Bioinformatics Center, College of Biology, Hunan University, Changsha 410082, China

[1]Co-first author

[*]Corresponding author: Zhongyang Tan Email: zhongyangtan@yeah.net

Abstract

Simple sequence repeats (SSRs) are found ubiquitously in almost all genome, and their formation mechanism is ambiguous yet. Here, the SSRs were analyzed in 55 randomly selected segments of genomes from a fairly wide range of species, with introducing more open standard for extensively mining repeats. A high percentage of repeats were discovered in these segments through that open standard and verified that they are not random. However, the current theory suggested that repeats tend to disappear over long-term evolution, which is inconsistent with such high proportion of SSRs remained in the genomes, inferring that there is most probably a mechanism for continually producing repeats during replicating process to balance continuous repeat disappearance. Therefore, the accepted straight-line slippage model is necessary to improve for explaining occurrence of simple sequence repeats. We proposed a folded slippage model to explain the high percent microsatellite occurrence with considering the geometric space of nucleotides and hydrogen bond stability, which can describe microsatellite expansion and contraction more reasonably. And analysis of external forces in the folding template strands showed the microsatellites tend to expand than contract. Our research may provide implements for the contribution of microsatellites to genome evolution and complement the semi-conservative replication.

Introduction

Simple sequence repeats (SSRs), also referred as microsatellites, have attracted increasingly great interests in recent decades [1-6], and have been widely analyzed in the genome sequences of eukaryotic prokaryotic and also viral genomes [2,5-7]. SSRs are the most variable genomic sequences,

which tend to appear frequent variations in repeat-unit number instead of nucleotide substitution. And it may be a critical power accelerate the genomic evolution [5,8], have roles associate with the host-adaptation and pathogenicity [8,9], be relevant with the expression of genes and activity of promoters [4,10], have relationship with many genetic diseases [11-13], and be observed with microsatellite instability (MSI) in many type of cancers [14-17].

Though SSRs have been comprehensively researched, there is actually no precise definition or wide-convinced standard for the extraction of SSRs all the time, which is usually based on setting the minimum numbers of the iterations for the mononucleotide to hexanucleotide SSRs based on empirical criterion [2,3,5,8]. Majority of previous studies showed more interesting into the relatively longer repetitive sequences [18-20], and most studies usually used the threshold of 6, 3, 3, 3, 3, 3 for extracting mono- to hexanuleotide SSRs [21-24], while the very short repeat-motifs with smaller iterations were almost excluded, causing the neglect of their important significance [25-28]. In this work, the selected SSRs were extensively extracted with a wider extracting standard for extensive repeat-motif grabbing to investigate the essential occurrences of SSRs.

It is widely accepted that DNA slippage is thought to be the primary mechanism for driving microsatellites expansion or contraction, however, slippage involves DNA polymerase pausing, dissociation and re-association [5,29,30], 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 [31-34]. It was suggested that the SSRs 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 [35,36]. 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 [37-40]. 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 SSRs. 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 microsatellite expansion and contraction. This work may also put forward some constructive suggestions for complementing the theory of semi-conservative replication.

Results and discussion

Genomes tend to produce short repeats

We analyzed 55 randomly-selected reported segment sequences covering from animal, plant, fungus, protist, bacteria, archaea and viruses (Supplementary Table 1). The SSRs were extracted from all these segment sequences by 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 [2,5,25,26,28], we found they occurred in a very large number. It is difficult to consider them just as random sequences but not repetitive sequences, and it is also inappropriate to consider the iteration of 3 to 5 of mononucleotide repeats just as random sequences. Therefore, the threshold was set at 3, 2, 2, 2, 2, 2 in this study for exploring more comprehensive occurrence of SSRs, which could grab shorter simple repeats that never analyzed before, and another two thresholds were used to analyze these sequences for comparison. To test whether the SSRs under this threshold are random, we generated 55 mimic sequences with same size and nucleotide composition to the corresponding 55 reported sequences.

The analyzed data showed that the reported segment sequences are averagely 44.4% constituted, with SSRs, ranging from 36.4% to 60.0% under the new threshold (Fig. 1a, Supplementary Table 1). And comparing analysis also show the SSR content of these segments with average of 18.8% and 5.0%. These results indicate that all these segments remained high content of SSRs, because all these segments are randomly selected from their genomes, suggesting that the remaining high content of short SSRs is a general feature of all organism genomes after long time evolution, and also

suggesting that few formerly well-studied repeats may only stand for the proverbial tip of the iceberg [2,3,5-7]. The null hypothesis test demonstrated that the percentages of SSRs in the generated segments are all lower than those in the reported segments, indicating that the high percentages of short SSRs are not randomly remained in all reported segments.

Though the evolutionary mechanism of nucleotide sequences is still hotly debated by evolutionist, it is widely accepted that the genomic sequences are continually mutating forever; and the neutral molecular evolution and molecular clock theory suggested that the nucleotide substitution is constant over the evolution time; the thermodynamics in biology states that an isolated system will always tend to disorder [41-46]. As the microsatellites are indeed ordered sequences, according to the former stated theories, the ordered repeats possibly tend to mutate into disordered sequences in the long evolutionary history without any selective pressure. Therefore, the repeat sequences should tend to disappear in genomes in the long evolution history. However, the remaining high percentage of SSRs in genomes is contradicted with the ideas of repetitive sequences tend to become no repetitive sequences. Thus, it can be inferred that there is most probably a mechanism for continually produce repeats to balance continuous repeat disappearance, and be responsible for the remaining of high percentage of short repeat sequences in genomes (Fig. 1b).

Furthermore, the SSRs of small iteration numbers were observed to occur largely more than those of large iteration numbers in all analyzed segments (Table 1, Supplementary Table 2), and this observation indicated that the SSRs of small iteration numbers maybe the basis for forming the SSRs of large iteration numbers, otherwise, it should be that the SSRs of large iteration numbers possibly are remained in higher percent level than or at least almost same level to the SSRs of small iteration numbers. Some of the longer SSRs also possibly mutate into short SSRs by contraction and point mutation as debated by many evolutionists [5,12,47], and these debates are possible because of that most of short repeats were not considered in their statistics; our observations generally suggested that most of the longer SSRs possibly evolved from short SSRs by expansion. So, the genomes possibly tend to produce short repeats by a continual repeat producing mechanism with the possibility of expansion a little more than that of contraction.

Relatively semi-conservative replication

It is well known that each base pair of DNA is one-to-one correspondence without other extra residue during replication in the double-helix model [35,36]. And Meselson and Stahl have verified the replication of DNA chains is semi-conservative by the sedimentation techniques based on the diversity differential of DNA with different isotopes, also implicating that the number of nucleotides in replicating strand is consistent with that in template strand while processing complete replication [48]. However, if the remained high percentage of short repeats is produced during replication process as described above, it certainly makes the base numbers of replication strand to be unequal to those of template strand, with one or several nucleotides/motifs being repeated and more than that in template strand. In vitro experiments also revealed the presence of repeats during DNA replication, and the nascent replication chain has a base increase [27,37,38,49]. And in this case, the replication process is possibly relatively semi-conservative and could be described as the following formula:

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

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

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 for the number of nucleotides 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 with relating to the possibility of repeat sequence 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 replicating strand is usually detected to be exactly equal to that in template strand, which is possibly because of the observed template strand being too short, for example, the total number of nucleotides in the initial template strand for stable PCR is up to two to three thousand nucleotides, in this case, we suppose $N_0 = 3000$, $\lambda_1 = 10^{-5}$, $f_1 = 1$, then the value of ΔN_1 will 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 template strand, and the discovery of new-born repeat is unavailable;

however, when the observed strand is long enough, then ΔN_i is able to be larger than 1 at least, and it can be found that the number of nucleotides in replicating strand is different from that in template strand, for instance, we suppose $N_0 = 10^6$, $\lambda_1 = 10^{-5}$, $f_1 = 1$, then the value of ΔN_1 will be 10, in this case, the replicating strands probably have 10 nucleotides (or repeat-motifs) more than template strand do after this replication. Thus, the increased number of nucleotides may represent newly occurred repeat sequences.

The occurrence of SSRs will possibly encounter selective pressure, though it may be different in coding or non-coding regions, then, we use f_i representing the fixation possibility of the newly born repeats facing with the selective pressure. The $f_i = 0$ when the occurrences of new repeats are the lethal mutations and unable fixation in the organism, or may be excluded by DNA repair system [1,50]. The fixation coefficient is $0<f_i<1$ when the new SSRs are the deleterious but fixed in the genome within alive individuals, like Huntington's disease [13]. While the occurrences of new SSRs are the neutral mutations, the fixation coefficient should be $0 \leq f_i \leq 1$, and they are fixed or excluded depending on genetic drift. And the f_i of beneficial mutations is 1, representing that the new SSRs may help the organism surviving. Therefore, the remaining high percentage of short repeats suggests that the replicating process possibly produce short repeat sequences frequently which may be fixed neutrally, beneficially, or deleteriously with diseases, and also suggests that the replication may be relatively semi-conservative.

Folded slippage model

The nucleotide chains of various species tend to produce simple repeats during replication, and thus cause the number of the nucleotides in replication strand possible to be different from template strand after replication as discussed above. Moreover, how did simple repeats actually originate from is still a key argument topic [5,47,51]. The widely accepted mechanism of occurring SSRs is the replication slippage model, which is possibly easy to explain the expansion and contraction of longer SSRs, but possibly difficult to explain the much amounts of short repeats expansion and contraction. And the current slippage model is indeed a straight template strand model, without considering that the space is required for nucleotide base and also phosphodiester bonds are much stronger than hydrogen bond (Fig. 2a) [52,53], and also without considering what is the force to drive the replicate

strand slippage. The straight replication slippage model has not given any clear suggestion, and it suggests that the SSRs possibly occurred by slippage occasionally [12,54-56]. Actually, there are about 33 atoms in a nucleotide (A: 33, T: 33, G: 34, C: 31) [57], and of course the nucleotide base need a certain space in nature. 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) [52,53,58]. We reconstructed the linear replication slippage model with a CAD geometric calculation by considering the space of bases (Fig. 2b, Supplementary Fig. 1); if the slippage bubble has enough geometric space to accommodate the repeat bases, the phosphodiester bond should be elongated far more than 0.34 nm, while the phosphodiester bonds in DNA is actually much stronger than hydrogen bond (Fig. 2a)[58]. So it is impossible to form a slippage bubble by a larger elongation of the phosphodiester bonds for accommodating the repeat bases. Therefore, the straight slippage model is very difficult to the occurrence of short repeats, and it is most possibly necessary to improve the slippage model.

Actually there is a fact which is widely ignored in replication slippage studies. The template strands are thought to be straight in all replication models, though it is the truth in general condition. It is also well known that the genomic DNA chains are very long and the space is too narrow in 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 [57]; therefore, the genomic DNA chains are generally highly curved and folded in the nucleus as widely accepted. Indeed, the replicating molecule is believed to be a straight molecule [37-40], and the replicating enzyme complexes usually straighten the template strand to be straight making the replicating strand well paired to finish the semi-conservative replication process [37,59,60]. However, there are a lot of environmental factors like temperature, viral proteins or diseases etc., which may disturb the normal works of the enzyme complexes. So, when the replicating enzyme complexes are disturbed by environmental factors, the replicating part DNA molecule may recover to some extent of curved or folded state, and then the template strand may also be some extent of curved or folded state.

Firstly, we proposed a curved template slippage model. When the curved DNA strand is used as

the template strand on inner side, the replication strand is longer than the template strand and can form more nucleotides than the template strand on the outside for during replication process. The replication strand should be longer than template strand, then, is able to provide extra spaces for accommodating the extra repeat bases (Fig. 3b). However, it is well known that the links of base pairs mainly depend on 2 types of hydrogen bonds, N—H … :N and N—H … :O [52], and the strengths of these hydrogen bonds are negatively correlated to the distance between every base pair; the strength of the hydrogen bond is about 3% of the 3', 5'-phosphodiester bonds [53,58,61,62] (Fig. 2a), so the distance between the bases is fixed; even if there is space to form a slippage bubble, the hydrogen bond should be elongated to exceed the threshold of 0.167 nm [52] and should be easy to be broken off in such condition. So, the curved slippage model is able to provide spaces for forming slippage bubble with forming unstable hydrogen bonds double-chain structures (Arm1 and Arm2) at both sides of the slippage bubble (Fig. 3b, Supplementary Fig. 2), indicating that the curved slippage model should be unreasonable.

Then we proposed a folded slippage model. In this case, the folded template strand forms a slippage bubble above the folding site to have sufficient space for accommodating the repeat nucleotides in replication process, the phosphodiester bonds are not elongated, but the bases are well paired with the stable hydrogen bonds at both sides of the slippage bubble (Fig. 4). If folding angle is proper, thereby it is most possibly to form a very stable double-stranded folded slippage structure to provide chances for producing repeats, with considering nucleotide geometric spaces and stability of phosphodiester and hydrogen bonds. Actually, there are two conditions of the folded slippage models: When template strand is on the inner side, the repeat unit duplicated to produce new repetitive unit or repeat expansion (Fig. 4); and when the template strand is on the outside, the replication strand may make the repetitive sequences to contract (Fig. 5); the features of this folded slippage model can easily explain the widely observed microsatellite mutations with expansion and contraction of repeat units [5,12,47,55,63]. 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 rotation angel of 18°, 36°, 54°, 72°, 90° and 108°, the replication strands will produce mononucleotide to hexanucleotide repeat expanding respectively (Fig. 4). So, it is

necessary to break off the number of hydrogen bonds from 2 to 18 without elongating the phosphodiester bond to produce repeats; it suggested that the difficulty of formation repeats from mono- to hexanucleotide is gradually increasing, and also means the occurrence of mono-, di-, tri-, tetra-, penta- and hexanucleotide repeat is gradually decreasing; that is well consistent with our statistic data (Table 1, Supplementary Table 2). Vice versa, when template chains are folded on the outside at a rotation angel of 18°, 36°, 54°, 72°, 90° and 108°, the replication strands will produce responding repeats contracting respectively (Fig. 5). These features are well corresponding to the microsatellites which usually refers to the tandem repeats with repeat units from mono- to hexanucleotides [2,5,6,24]. According to this rule, we also describe the possible folded template slippage models of hepta-, otca-, nona- and decanucleotide repeats (Supplementary Figs. 3 and 4). In fact, the replicating strand must break off at least from 14 to 30 hydrogen bonds to make a folded slippage bubble, the energy to break off so much hydrogen bonds are almost close to energy of phosphodiester bond, then, they are very difficult to occur, and therefore, this is consistent with the observations that such long tandem repetitive sequences are often not very abundant in the genomes [55,64]. The (A_mT_n) repeats growing faster than (G_mC_n) repeats also suggested that the broken number of hydrogen bonds involves in the speed of repeat expansion [18,34,65,66]. Although this folded slippage model is just simply described in a plane form, it can still clearly simulate and explain the repeat sequences producing process. We also use the same space size to make the double-helical three-dimensional forms show the folded slippage model more intuitively (Figs. 4 and 5), and the precise folding angle in the three-dimensionally double-helical forms and other issues desire further study.

There is enough geometric space in the slippage bubble of the folded template model to accommodate repeat nucleotides without stretching the phosphodiester bonds, compared with the straight template slippage model. In contrast to the curved template model, the difference in the folded model is that the two sides of the slippage bubble are stably paired, and the Arm1 and Arm2 similar to the straight template replication model are formed 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, and is easy used to explain microsatellite mutations with repeat unit expansion and contraction. Therefore, we propose

that the folded template chain slippage model may be considered as the most reasonable model for explaining repeats production in replicating process, and the folded template strand slippage model may be responsible for the continual producing of repeat sequences and the remaining of high percentage of repeat sequences in genomes.

Microsatellites tend to expand

As stated above, according to the folded slippage model, template chain folding on the inner side may make the replicating chain slippage for repeats expansion, vice versa, the template chain folding on the outside may make the replicating chain slippage for repeats contraction; and it seems that the possibility of repeats expansion and contraction is same. However, there are two manners for the repeat sequences contraction, one is above mentioned the template chain folds on outside, another is also above stated general mutations; the high content of the repeat sequence is still in a stable state in the genome of each species, suggesting that the possibility of repeat expansion should be higher than repeat contraction. And many reports also suggest that there is a higher possibility of repeat expansion than repeat contraction [27,49,67].

When the folded template chain slippage was deeply investigated, the replicating straight template DNA chain should return to folded under external forces from the narrow and crowded cell nucleus when the replicating enzyme complexes are disturbed, and usually the replicating enzyme complexes may provide power for balancing the external forces to drag the template DNA molecule straight. Then, we proposed an external force function for template strand returning to folded, and this function may be helpful to explore the probability of expansion and contraction. When the template strand is on the inner side, the nucleotide bases are outward, and the space of bases at the folded site become wide and loose at outward part; while it is on the outside, the base in the folding position is squeezed inward. Comprehensively considering the small difference of the space of nucleotides at the folded site, it can be easy accepted that the external forces to make template strand folded with bases loose should be smaller than that to be squeezed; therefore, the external force required for the template strand folded on the outside (F^o) is inevitable greater than that (F^i) on the inner side, it can be described as $F^o > F^i$, suggesting that the probability for the template strand folded on the inner side is higher than that on the outside; as our folded slippage model suggested that the repeats tend to

expand when the template strand on inner side and contract when the template strand on outside, therefore, the possibility of repeat expansion (P^e) is most possibly higher than that for repeat contraction (P^c), it can be described as $P^e > P^c$ (Fig. 6). The SSR studies, like in Huntington disease related locus and myotonic dystrophy type 1 locus, all showed SSR expansion biased [12,13,68-70], which proving that the expansion and of short SSRs are more frequent than that of contraction.

Thus, according to formula (2):

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

thus, $\Delta N^c = N^c_i - N^c_{i-1} = \mathrm{int}[N_0 f^c_i \lambda^c_i (1+ f^c_1 \lambda^c_1)(1+ f^c_2 \lambda^c_2)\ldots(1+f^c_{i-1}\lambda^c_{i-1})] \leq 0$.

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

thus, $\Delta N^e = N^e_j - N^e_{j-1} = \mathrm{int}[N_0 f^e_j \lambda^e_j (1+ f^e_1 \lambda^e_1)(1+ f^e_2 \lambda^e_2)\ldots(1+f^e_{j-1}\lambda^e_{j-1})] \geq 0$.

The general repeat expansion and contraction can be described as:

$$|\sum \Delta N^e| = |\mathrm{int}[\sum N_0 f^e_j \lambda^e_j (1+ f^e_1 \lambda^e_1)(1+ f^e_2 \lambda^e_2)\ldots(1+f^e_{j-1}\lambda^e_{j-1})]|;$$

$$|\sum \Delta N^c| = |\mathrm{int}[\sum N_0 f^c_i \lambda^c_i (1+ f^c_1 \lambda^c_1)(1+ f^c_2 \lambda^c_2)\ldots(1+f^c_{i-1}\lambda^c_{i-1})]|;$$

$\sum \Delta N = |\sum \Delta N^e| - |\sum \Delta N^c| = \mathrm{int}[N_0 \sum[|f^e_j \lambda^e_j (1+ f^e_1 \lambda^e_1)(1+ f^e_2 \lambda^e_2)\ldots(1+f^e_{j-1}\lambda^e_{j-1})| - |f^c_i \lambda^c_i (1+f^c_1 \lambda^c_1)(1+f^c_2 \lambda^c_2)\ldots (1+f\lambda^c_{i-1}\lambda^c_{i-1})|]]$.

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 ($|\lambda^c|$), if we suppose that $f^e = f^c = f$, $i = j$, and as generally $P^e > P^c$, then $\lambda^e > |\lambda^c|$,

and also $\sum[|\lambda^e_j (1+ f\lambda^e_1)(1+ f\lambda^e_2)\ldots(1+f\lambda^e_{j-1})|] \geq \sum[|\lambda^c_i (1+f\lambda^c_1)(1+f\lambda^c_2)\ldots (1+f\lambda^c_{i-1})|]$,

therefore, $\sum \Delta N = |\sum \Delta N^e| - |\sum \Delta N^c| \geq 0$.

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 remaining of high percentage of short repeats in genomes under a mechanism of continually producing repeats; and this mechanism might result from the folded template chain slippage model, which is possibly responsible for the widely occurring short tandem repeats, also called microsatellites or SSRs in eukaryotic, prokaryotic and also viral genomes. We improved the straight slippage model to folded slippage model by fully considering the geometric spaces of nucleotides base, the relationship between phosphodiester and hydrogen bond and the

stability of these bonds. The slippage model showed that the straight replicating template DNA may return to be some extent of folded 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 related with the former slippage model [30,55]. The easily forming of folded slippage may be also responsible for the widely observed fact that repetitive part of genome is usually evolved hundred or more times than other part with only repeat units expansion and contraction [1,17,47,71], though the repeats occurred more in non-coding regions than in coding regions possibly because of different selective pressures [5,12,55]. Most of new occurring repeats should be lethal mutation and may have been negatively selected to lost; some of new occurring repeats should be deleterious in genomes and responsible for a series of diseases [69,70,72,73]; many neutral repeat expansions may be lost or fixed with no functions in genomes by genetic drift [74]; and some beneficial repeat expansions may promote the emergence of different new properties or functions, that is why the repeat sequences are reported with so many different roles [8-10,63,65,75,76]. And the longer repeats might originate from short repeat expansion by the folded template slippage, and the longer genomes possibly evolved from the short genome with related to the continuous repeats producing folded slippage model in the long evolutionary replicating process.

References

1. Mandal, R. et al. CANCER Genetic diversity of tumors with mismatch repair deficiency influences anti-PD-1 immunotherapy response. *Science* **364**, 485-491 (2019).
2. Zhao, X. et al. Coevolution between simple sequence repeats (SSRs) and virus genome size. *Bmc Genomics* **13**, 435-435 (2012).
3. Chen, M., Tan, Z., Zeng, G. & Peng, J. Comprehensive analysis of simple sequence repeats in pre-miRNAs. *Molecular Biology and Evolution* **27**, 2227-2232 (2010).
4. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K.J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213-1236 (2009).
5. Ellegren, H. Microsatellites: Simple sequences with complex evolution. *Nature Reviews: Genetics* **5**, 435-445 (2004).
6. Morgante, M., Hanafey, M. & Powell, W. Microsatellites are preferentially associated with

nonrepetitive DNA in plant genomes. *Nature Genetics* **30**, 194-200 (2002).

7. Lin, W.H. & Kussell, E. Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. *Nucleic Acids Research* **40**, 2399-2413 (2012).
8. Li, Y.C., Korol, A.B., Fahima, T. & Nevo, E. Microsatellites Within Genes: Structure, Function, and Evolution. *Molecular Biology and Evolution* **21**, 991-1007 (2004).
9. Hood, D.W. et al. DNA repeats identify novel virulence genes in Haemophilus influenzae. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 11121-11125 (1996).
10. Hannan, A.J. Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews: Genetics* **19**, 286-298 (2018).
11. Jain, A. & Vale, R.D. RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243-247 (2017).
12. Mirkin, S.M. Expandable DNA repeats and human disease. *Nature* **447**, 932-940 (2007).
13. Macdonald, M.E. et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**, 971-983 (1993).
14. Chan, E.M. et al. WRN helicase is a synthetic lethal target in microsatellite unstable cancers. *Nature* **568**, 551-556 (2019).
15. Bailey, M.H. et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e18 (2018).
16. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews: Genetics* **15**, 585-598 (2014).
17. Kim, T.M., Laird, P.W. & Park, P.J. The Landscape of Microsatellite Instability in Colorectal and Endometrial Cancer Genomes. *Cell* **155**, 858-868 (2013).
18. Tian, X.J., Strassmann, J.E. & Queller, D.C. Genome Nucleotide Composition Shapes Variation in Simple Sequence Repeats. *Molecular Biology And Evolution* **28**, 899-909 (2011).
19. Kelkar, Y.D. et al. What Is a Microsatellite: A Computational and Experimental Definition Based upon Repeat Mutational Behavior at A/T and GT/AC Repeats. *Genome Biology And*

Evolution **2**, 620-635 (2010).

20. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580 (1999).
21. Chen, M., Tan, Z. & Zeng, G. Microsatellite is an important component of complete hepatitis C virus genomes. *Infection, Genetics and Evolution* **11**, 1646-1654 (2011).
22. Rajendrakumar, P., Biswal, A.K., Balachandran, S.M., Srinivasarao, K. & Sundaram, R.M. Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. *Bioinformatics* **23**, 1-4 (2007).
23. George, B., Alam, C.M., Jain, S.K., Sharfuddin, C. & Chakraborty, S. Differential distribution and occurrence of simple sequence repeats in diverse geminivirus genomes. *Virus Genes* **45**, 556-566 (2012).
24. Zhao, X. et al. Microsatellites in different Potyvirus genomes: survey and analysis. *Gene* **488**, 52-56 (2011).
25. Teh, B.T. et al. The draft genome of tropical fruit durian (Durio zibethinus). *Nature Genetics* **49**, 1633-1641 (2017).
26. Hunt, V.L. et al. The genomic basis of parasitism in the Strongyloides clade of nematodes. *Nature Genetics* **48**, 299-307 (2016).
27. Fungtammasan, A. et al. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Research* **25**, 736-749 (2015).
28. Schmutz, J. et al. A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics* **46**, 707-713 (2014).
29. Gadgil, R., Barthelemy, J., Lewis, T. & Leffak, M. Replication stalling and DNA microsatellite instability. *Biophysical Chemistry* **225**, 38-48 (2016).
30. Viguera, E., Canceill, D. & Ehrlich, S.D. Replication slippage involves DNA polymerase pausing and dissociation. *The EMBO journal* **20**, 2587-2595 (2001).
31. Huang, T.Y. et al. Parity-dependent hairpin configurations of repetitive DNA sequence promote slippage associated with DNA expansion. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 9535-9540 (2017).

32. Garcia-Diaz, M., Bebenek, K., Krahn, J.M., Pedersen, L.C. & Kunkel, T.A. Structural analysis of strand misalignment during DNA synthesis by a human DNA polymerase. *Cell* **124**, 331-342 (2006).

33. Lai, Y.L. & Sun, F.Z. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular Biology and Evolution* **20**, 2123-2131 (2003).

34. Schlötterer, C. & Tautz, D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Research* **20**, 211-215 (1992).

35. Watson, J.D. & Crick, F.H.C. Molecular structure of deoxypentose nucleic acids. *Nature* **171**, 738-740 (1953).

36. Watson, J.D. & Crick, F.H.C. Genetical implications of the structure of dexoyribonucleic acid. *Nature* **171**, 964-967 (1953).

37. Kiefer, J.R., Mao, C., Braman, J.C. & Beese, L.S. Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. *Nature* **391**, 304 (1998).

38. Doublié, S., Tabor, S., Long, A.M., Richardson, C.C. & Ellenberger, T. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 A resolution. *Nature* **391**, 251-258 (1998).

39. Bell, S.D. DNA replication: archaeal oriGINS. *Bmc Biology* **9**, 36 (2011).

40. Costa, A. et al. The structural basis for MCM2-7 helicase activation by GINS and Cdc45. *Nature Structural & Molecular Biology* **18**, 471-477 (2011).

41. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275-6 (1977).

42. Kimura, M. The neutral theory of molecular evolution. *Scientific American* **241**, 98-100, 102, 108 passim (1979).

43. Margoliash, E. Primary Structure And Evolution Of Cytochrome C. *Proceedings of the National Academy of Sciences of the United States of America* **50**, 672-9 (1963).

44. Zuckerkandl, E. & Pauling, L.B. Molecular disease, evolution, and genic heterogeneity. in *Horizons in Biochemistry* (eds. Pullman, B., Kasha, M. & SzentGyo□rgyi, A.) 189-225 (Academic Press, New York, New York, 1962).

45. Zuckerkandl, E. & Pauling, L.B. Evolutionary divergence and convergence in proteins. in

Evolving Genes and Proteins (eds. Bryson, V. & Vogel, H.J.) 97-166 (Academic Press, New York, New York, 1965).

46. Bharadwaj, S., Montazeri, R. & Haynie, D.T. Direct determination of the thermodynamics of polyelectrolyte complexation and implications thereof for electrostatic layer-by-layer assembly of multilayer films. *Langmuir* **22**, 6093-6101 (2006).

47. Kelkar, Y.D., Eckert, K.A., Chiaromonte, F. & Makova, K.D. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Research* **21**, 2038-2048 (2011).

48. Meselson, M. & Stahl, F.W. The Replication of DNA in Escherichia Coli. *Proceedings of the National Academy of Sciences of the United States of America* **44**, 671-682 (1958).

49. Fungtammasan, A. et al. Reverse Transcription Errors and RNA-DNA Differences at Short Tandem Repeats. *Molecular Biology And Evolution* **33**, 2744-2758 (2016).

50. Jeggo, P.A., Pearl, L.H. & Carr, A.M. DNA repair, genome stability and cancer: a historical perspective. *Nature Reviews: Cancer* **16**, 35 (2015).

51. Torresen, O.K. et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research* **47**, 10994-11006 (2019).

52. Heyrovska, R. Dependence of the length of the hydrogen bond on the covalent and cationic radii of hydrogen, and additivity of bonding distances. *Chemical Physics Letters* **432**, 348-351 (2006).

53. Gao, F., Yin, C. & Yang, P. Coordination chemistrymimics of nuclease-activity in the hydrolytic cleavage of phosphodiester bond. *Chinese Science Bulletin* **49**, 1667-1680 (2004).

54. Leclercq, S., Rivals, E. & Jarne, P. DNA Slippage Occurs at Microsatellite Loci without Minimal Threshold Length in Humans: A Comparative Genomic Approach. *Genome Biology and Evolution* **2**, 325-335 (2010).

55. Gemayel, R., Vinces, M.D., Legendre, M. & Verstrepen, K.J. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annual Review of Genetics* **44**, 445-477 (2010).

56. Ohshima, K. & Wells, R.D. Hairpin formation during DNA synthesis primer realignment in vitro in triplet repeat sequences from human hereditary disease genes. *The Journal of Biological Chemistry* **272**, 16798-1806 (1997).

57. Alberts, B. et al. *Molecular Biology of the Cell (4th ed.)*, (Garland Science, New York, 2002).

58. Wang, Q. *Hydrogen Bond in Organic Chemistry*, (Tianjin University Press, Tianjin, China, 1993).

59. Costantino, L. et al. Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* **343**, 88-91 (2014).

60. Fragkos, M., Ganier, O., Coulombe, P. & Mechali, M. DNA replication origin activation in space and time. *Nature Reviews: Molocular Cell Biology* **16**, 360-374 (2015).

61. Luo, Y.R. *Comprehensive Handbook of Chemical Bond Energies*, (CRC Press, Boca Raton, FL, 2007).

62. Griffiths, A.J., Miller, J.H., Suzuki, D.T. & Lewontin, R.C. *An Introduction to Genetic Analysis, 7th edition*, (2000).

63. Gymrek, M. et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature Genetics* **48**, 22-29 (2016).

64. Legendre, M., Pochet, N., Pak, T. & Verstrepen, K.J. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research* **17**, 1787-1796 (2007).

65. Sinai, M.I.T. et al. AT-dinucleotide rich sequences drive fragile site formation. *Nucleic Acids Research* **47**, 9685-9695 (2019).

66. Katti, M.V., Ranjekar, P.K. & Gupta, V.S. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Molecular Biology And Evolution* **18**, 1161-1167 (2001).

67. Neil, A.J., Liang, M.U., Khristich, A.N., Shah, K.A. & Mirkin, S.M. RNA-DNA hybrids promote the expansion of Friedreich's ataxia (GAA)(n) repeats via break-induced replication. *Nucleic Acids Research* **46**, 3487-3497 (2018).

68. Higham, C.F., Morales, F., Cobbold, C.A., Haydon, D.T. & Monckton, D.G. High levels of somatic DNA diversity at the myotonic dystrophy type 1 locus are driven by ultra-frequent

expansion and contraction mutations. *Human Molecular Genetics* **21**, 2450-2463 (2012).

69. Sznajder, L.J. & Swanson, M.S. Short Tandem Repeat Expansions and RNA-Mediated Pathogenesis in Myotonic Dystrophy. *International Journal Of Molecular Sciences* **20**(2019).

70. Larson, E., Fyfe, I., Morton, A.J. & Monckton, D.G. Age-, tissue- and length-dependent bidirectional somatic CAG*CTG repeat instability in an allelic series of R6/2 Huntington disease mice. *Neurobiol Dis* **76**, 98-111 (2015).

71. Giesselmann, P. et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nature Biotechnology* **37**, 1478-1481 (2019).

72. Arturo, L.C., Cleary, J.D. & Pearson, C.E. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nature Reviews: Molecular Cell Biology* **11**, 165-170 (2010).

73. Sun, J.H. et al. Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* **175**, 224-238 (2018).

74. Muller, M.J., Neugeboren, B.I., Nelson, D.R. & Murray, A.W. Genetic drift opposes mutualism during spatial population expansion. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 1037-1042 (2014).

75. Mrazek, J. Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. *Mol Biol Evol* **23**, 1370-85 (2006).

76. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K.J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213-1216 (2009).

77. Lamond, A.I. Molecular Biology of the Cell, 4th edition. *Nature* **417**, 383-383 (2002).

Methods

Sequences resource

We downloaded 55 genomic sequences from Genbank of a fairly wide range of species that covering animals, plants, fungi, protozoa, bacteria, archaea and viruses. The segments for SSR analysis were randomly selected from different regions of these 55 genomic sequences, which range from 3000 to 96600 bp in length and do not contain any gaps, to verify the widespread distributions

of SSRs, as the full genomic sequences are too long.

Repeat extraction

The perfect simple sequence repeats were extracted by Imperfect Microsatellite Extraction Webserver (IMEx-web, http://imex.cdfd.org.in/IMEX/index.html) from those 55 randomly-selected reported 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 confined to ensure to be commonly recognized as the SSRs.

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 (Supplementary Program 1) according to the nucleotide compositions and sizes of those 55 reported segments. Then, the validating test, which can verify that the short SSRs extracted in those 55 reported segments are nonrandom sequences, was based on the comparison of the SSR percentages in the reported segments and our 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 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, which is an industrial drawing software.

Acknowledgments

The authors thank Qijun Tian, who is from School of design in Hunan University, for his technological help at designing the folded slippage model in helix structure.

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).

Author contributions

Author contributions: Z. Tan designed and directed this study. H. Zhang , D. Li and X. Zhao performed the data analysis, S. Pan and X. Wu helped for performing data analysis. H. Zhang and S. Pan drew the models, D. Li, S. Pan and X. Wu selected the segmental sequences. Z. Tan and H. Zhang prepared the manuscript. D. Li and S. Peng helped for the manuscript preparation.

Competing interests

The authors declare no competing interests.

Figure legends

Fig. 1 A high percentage of SSRs in genomes and genomes probably tend to produce repeats. (*A*) SSR percentages of 55 randomly-selected reported segments and the control group, which consisted of the generated segments according to the sizes and nucleotide compositions of corresponding reported segments. (*B*) Contradiction analysis of disappearance and high percentage of SSRs in the genomes.

Fig. 2 Straight strand models of semi-conservative replication and slippage. (*A*) The space of a nucleotides 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)[52,53,58]. # 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.

Fig. 3 The DNA molecule 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) [77]; The normal replicating enzymes complex straighten the DNA molecule, 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).

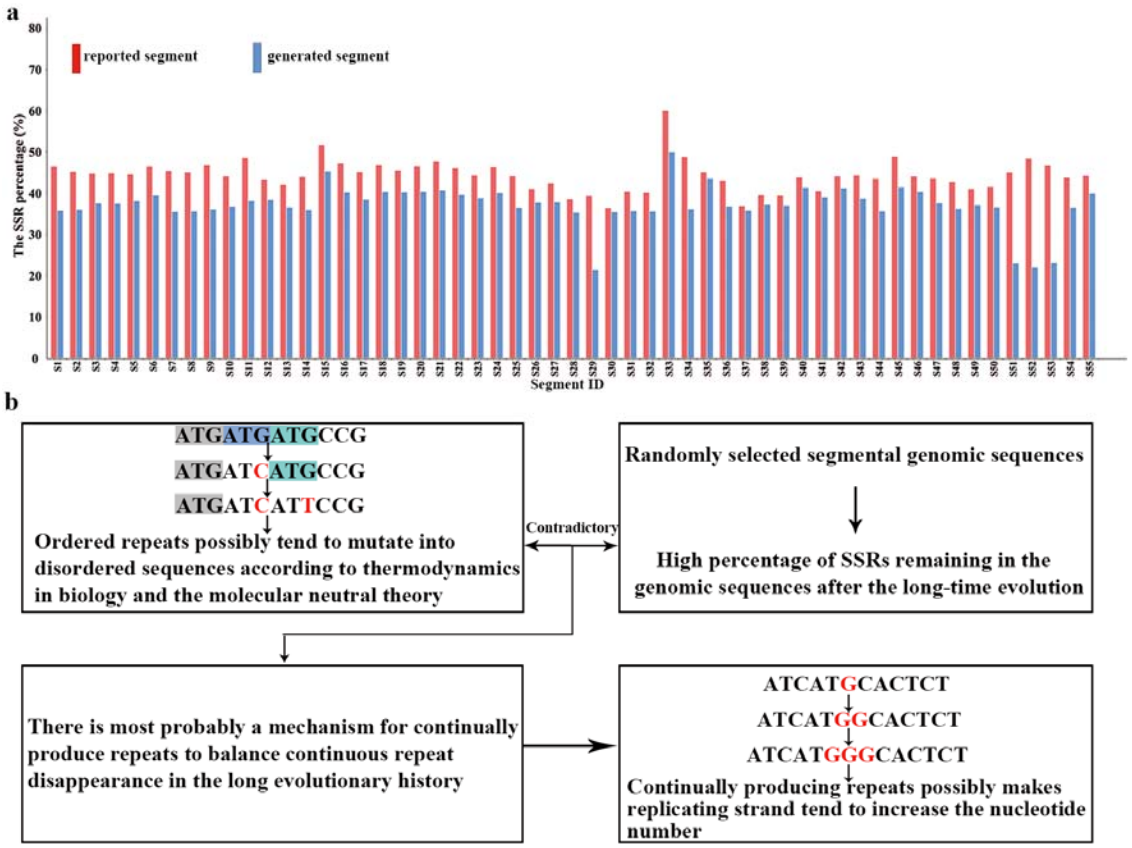
Fig. 4 Stable folded slippage models of mononucleotide to hexanucleotide repeats amplification according to the strict geometric calculation of the space of a nucleotide and the stability of hydrogen and phosphodiester bonds. Repeat units tend to be expanded in the replicating strands when the template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

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 figures were the folded slippage models in three-dimensional helix form.

Fig. 6 Repeats 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.

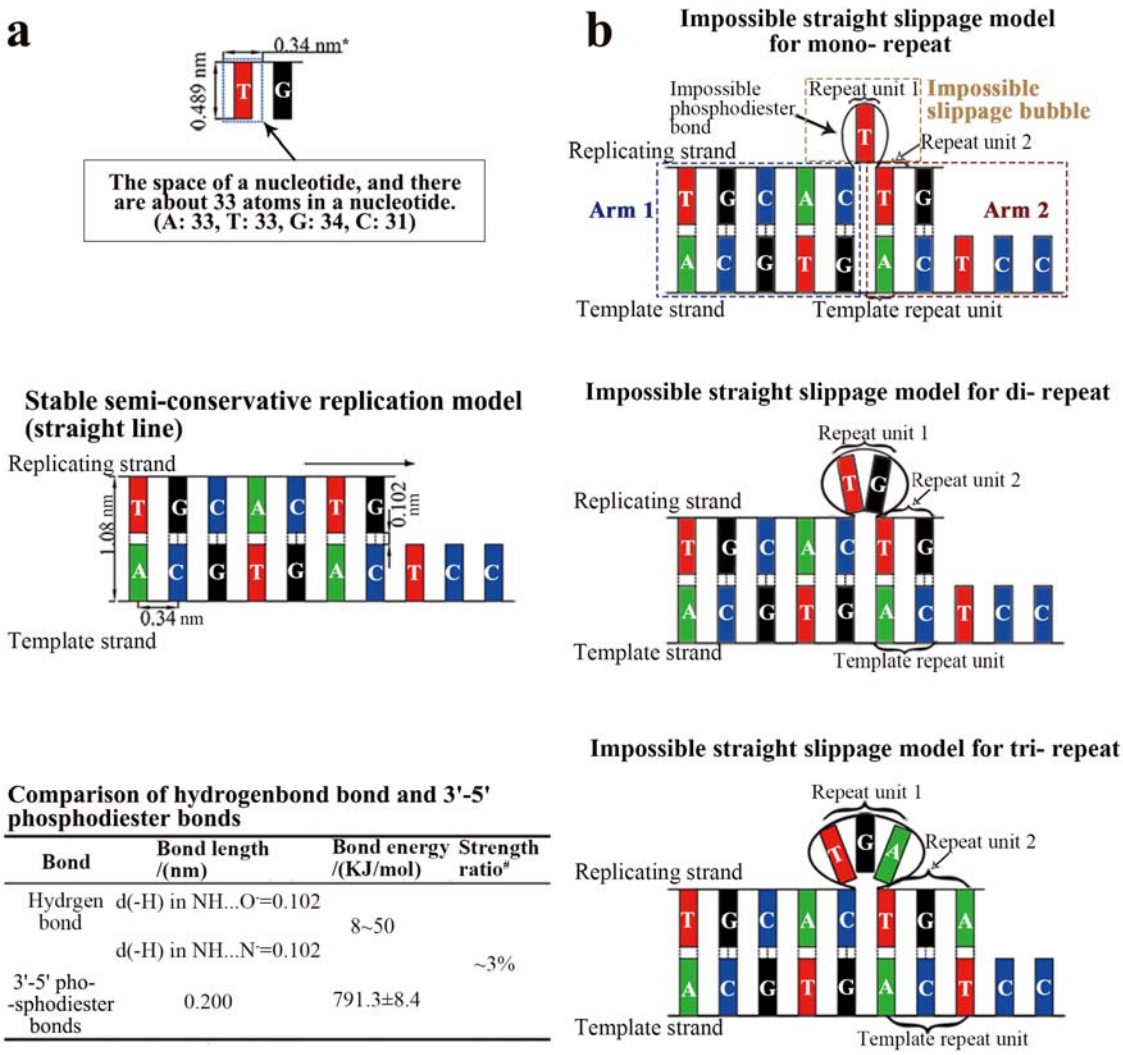
Figures

Fig. 1



A high percentage of SSRs in genomes and genomes probably tend to produce repeats. **a**, SSR percentages of 55 randomly-selected reported segments and the control group, which consisted of the generated segments according to the sizes and nucleotide compositions of corresponding reported segments. **b**, Contradiction analysis of disappearance and high percentage of SSRs in the genomes.

Fig. 2

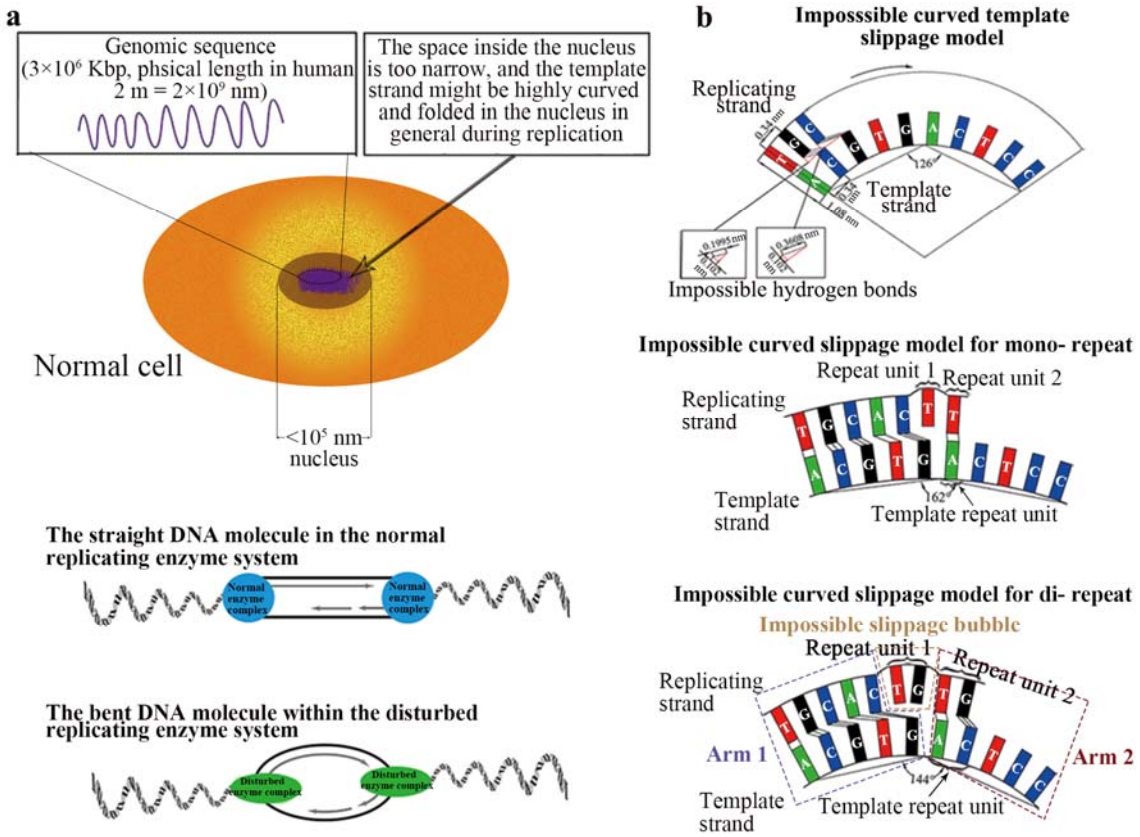


Comparison of hydrogenbond bond and 3'-5' phosphodiester bonds

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

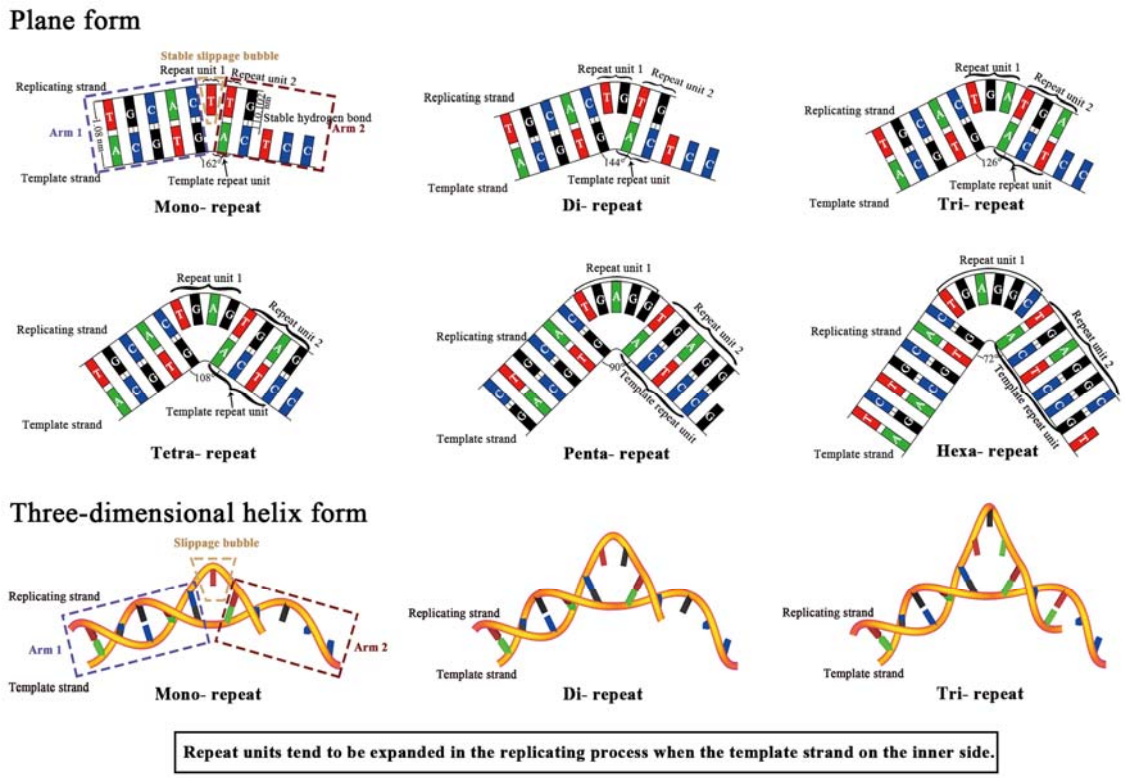
Straight strand models of semi-conservative replication and slippage. **a**, The space of a nucleotides was drawn. * indicates that those number is the theoretical values (top); The stable straight model of semi-conservative replication (middle); The comparison of hydrogen bond and 3'-5' phosphodiester bonds (bottom)[52,53,58]. # 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.

Fig. 3



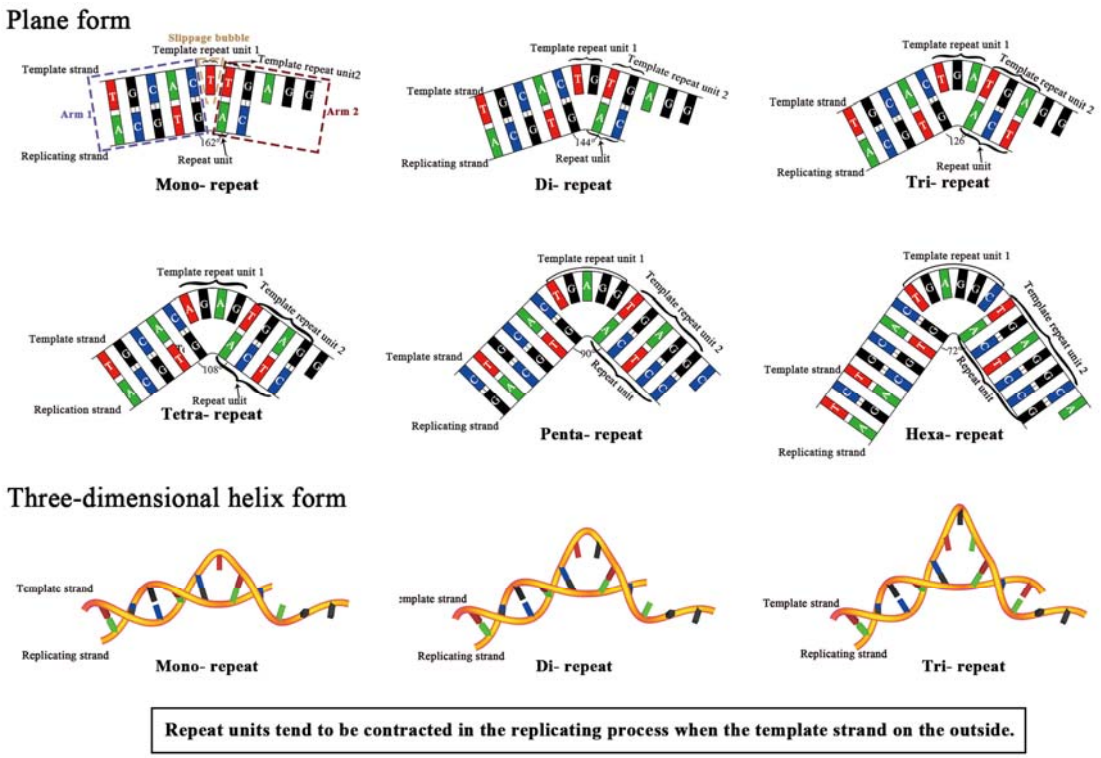
The DNA molecule is highly curved or folded in the nucleus and the impossible curved slippage model. **a**, Schematic diagram of the size of the nuclear space (top) [77]; The normal replicating enzymes complex straighten the DNA molecule, while the disturbed replicating enzymes complex may cause the 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).

Fig. 4



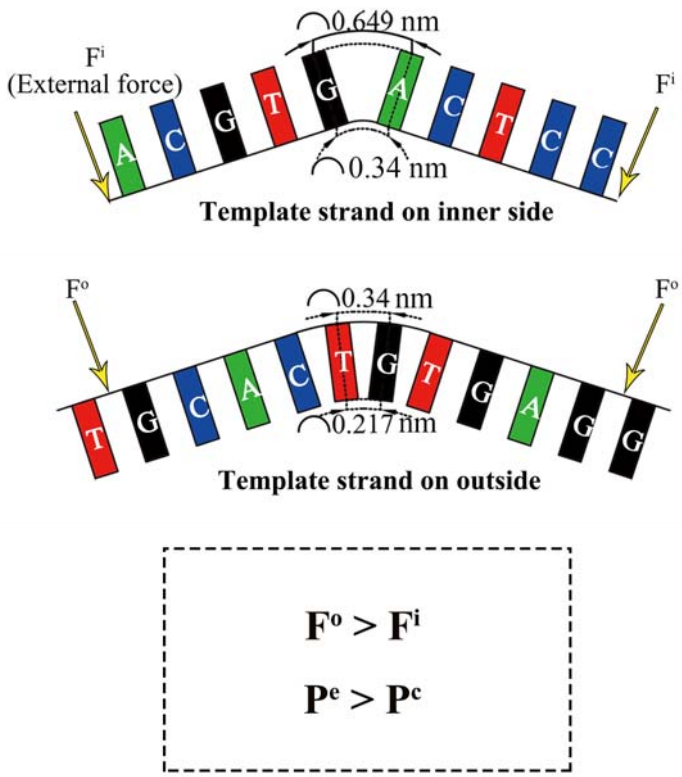
Stable folded slippage models of mononucleotide to hexanucleotide repeats amplification according to the strict geometric calculation of the space of a nucleotide and the stability of hydrogen and phosphodiester bonds. Repeat units tend to be expanded in the replicating strands when the template strands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

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 figures were the folded slippage models in three-dimensional helix form.

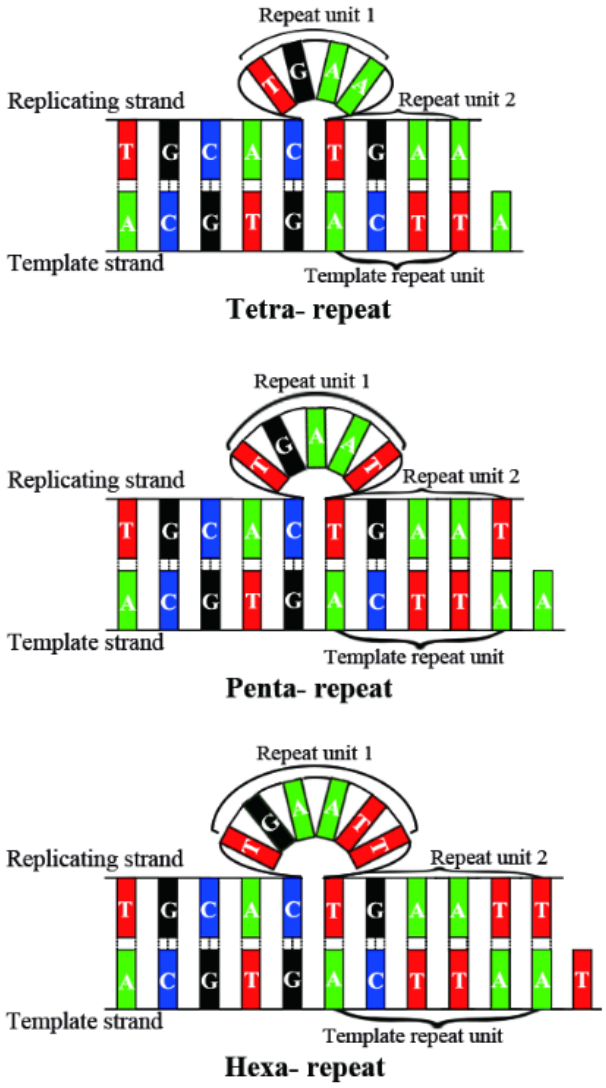
Fig. 6



Repeats 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.

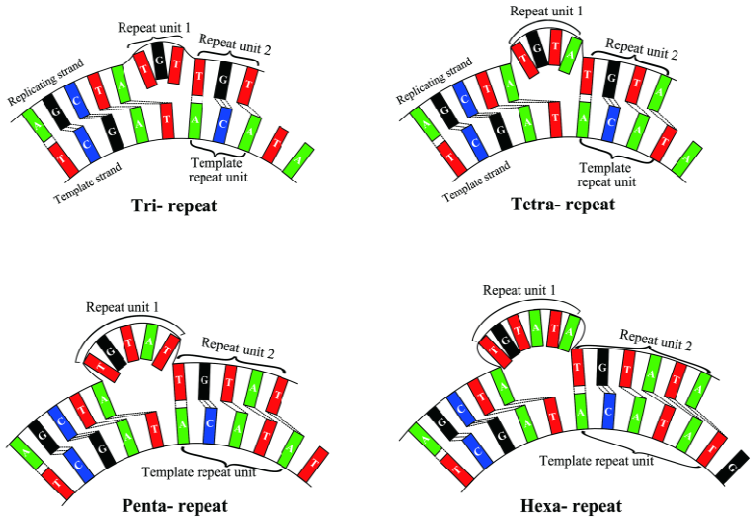
Supplementary Figures

Supplementary Fig. 1



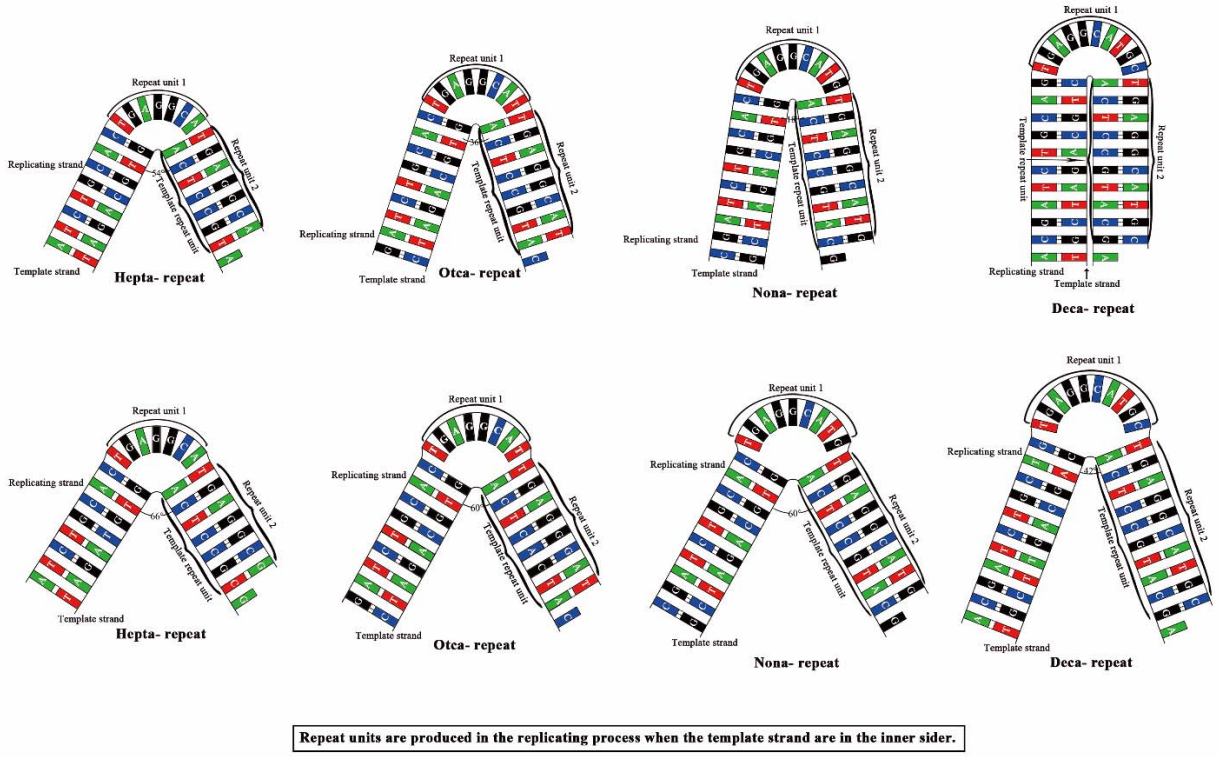
Impossible straight slippage models for tetra- to hexanucleotide repeats when the slippage bubble occurs at the replicating strand. The model drawing was based on the strict geometric calculation of the space of a nucleotide and the stability of hydrogen and phosphodiester bonds.

Supplementary Fig. 2



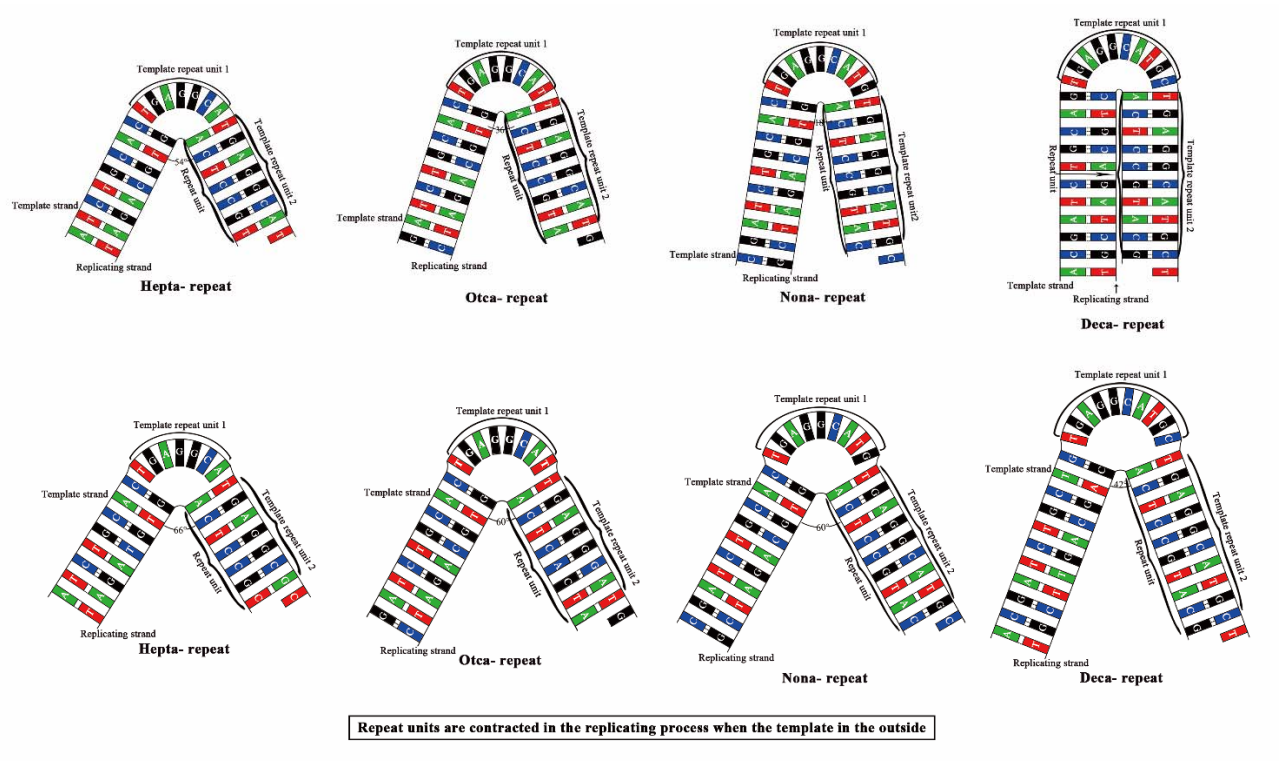
Impossible curved slippage models for tri- to hexanucleotide repeats when the template strand in the inner side of the models.

Supplementary Fig. 3



The possible folded slippage models for hepta- to decanucleotide repeat amplification. 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.

Supplementary Fig. 4



The possible folded slippage models for hepta- to decanucleotide repeat contraction. Repeat units tend to be subtracted in the replicating strands when the template strands are on the outside of the folded slippage models respectively.