Mini-review

Structures and conformational dynamics of DNA minidumbbells in pyrimidine-rich repeats associated with neurodegenerative diseases

Yuan Liu [a],[1], Liqi Wan [b],[1], Cheuk Kit Ngai [c],[1], Yang Wang [d], Sik Lok Lam [c], Pei Guo [d],[*]

[a] South China Advanced Institute for Soft Matter Science and Technology, School of Emergent Soft Matter, South China University of Technology, Guangzhou, Guangdong 510640, China
[b] School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[c] Department of Chemistry, The Chinese University of Hong Kong, Shatin N.T., Hong Kong, China
[d] Zhejiang Cancer Hospital, Institute of Basic Medicine and Cancer (IBMC) Chinese Academy of Sciences, Hangzhou, Zhejiang 310022, China

ARTICLE INFO

ABSTRACT

Expansions of short tandem repeats (STRs) are associated with approximately 50 human neurodegenerative diseases. These pathogenic STRs are prone to form non-B DNA structure, which has been considered as one of the causative factors for repeat expansions. Minidumbbell (MDB) is a relatively new type of non-B DNA structure formed by pyrimidine-rich STRs. An MDB is composed of two tetraloops or pentaloops, exhibiting a highly compact conformation with extensive loop-loop interactions. The MDB structures have been found to form in CCTG tetranucleotide repeats associated with myotonic dystrophy type 2, ATTCT pentanucleotide repeats associated with spinocerebellar ataxia type 10, and the recently discovered ATTTT/ATTTC repeats associated with spinocerebellar ataxia type 37 and familial adult myoclonic epilepsy. In this review, we first introduce the structures and conformational dynamics of MDBs with a focus on the high-resolution structural information determined by nuclear magnetic resonance spectroscopy. Then we discuss the effects of sequence context, chemical environment, and nucleobase modification on the structure and thermostability of MDBs. Finally, we provide perspectives on further explorations of sequence criteria and biological functions of MDBs.

© 2023 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Short tandem repeats (STRs) are DNA sequences consisting of 1–6-bp repeating units connected in a head-to-tail manner [1]. They are widely distributed in both the coding and non-coding regions and constitute ∼3% of the human genomic DNA [2]. STRs were originally thought as junk DNAs until their biological roles were recognized in early 1990 s [3,4]. In particular, genetic instability of STRs has been implicated in cellular dysfunctions associated with human diseases. Approximately 13 types of STRs, including CAG, CTG, CGG and GAA trinucleotide repeats, CCTG tetranucleotide repeats, TGGAA, ATTCT, TTTTA, TTTCA and AAGGG pentanucleotide repeats, and GGCCTG, CCCTCT and GGGGCC hexanucleotide repeats, have been well documented to exhibit dynamic mutations that change their sequence lengths during successive generations or individual's life [5–7]. These STRs interfere with cellular processes like DNA replication, transcription and translation when the number of repeats reaches a certain threshold, leading to some neurodegenerative disorders which are also called as repeat expansion diseases [8]. Nearly 50 repeat expansion diseases have been reported so far, such as the fragile X syndrome [9], Huntington's disease [10] and amyotrophic lateral sclerosis and frontotemporal dementia (ALS/FTD) [11]. For some of these diseases, a larger number of repeats in the mutated gene was found to correlate with a higher disease severity and an earlier disease onset [12–14].

Pathogenic mechanisms of repeat expansion diseases are very complex at various molecular levels including DNA, RNA and protein [15]. For repeat expansions occurring in the coding region, RNA transcripts harbouring expanded repeats are translated into extra-long proteins which lose normal functions [15]. There are nine polyglutamine (poly-Q) diseases caused by CAG repeat expansions in the coding region of respective genes, and the translated proteins containing long tracts of glutamines form aggregates that cause cellular toxicity [16]. For the remaining diseases associated with

---

* Corresponding author.
  E-mail address: guopei@ibmc.ac.cn (P. Guo).
  [1] Authors who contributed equally to this work.

non-coding repeat expansions, toxic RNA gain-of-function is involved in pathogenic mechanisms. Expanded RNA repeats form secondary structures and sequester various RNA-binding proteins to form nuclear foci, causing nuclear stress and mRNA splicing errors [17,18]. Besides, some expanded RNA repeats such as rGGCCUG and rGGGGCC are reported to produce polypeptide repeats through a non-AUG translation pathway [19,20].

Despite a complexity in the pathogenic pathways, expansions of STRs in the genomic DNA are origins of these neurodegenerative diseases. Therefore, a thorough understanding on the molecular mechanisms of repeat expansions lays the foundation for developing therapeutic approaches, especially those targeting at a DNA level. In 1995, Gacy et al. first reported the formation of hairpin structures in CAG and CTG repeats, and provided a strand slippage model to explain the occurrence of repeat expansions [21]. Since then, accumulating evidences have revealed the roles of non-B DNAs in causing repeat expansions and cellular dysfunctions [22]. The structures of STRs have been widely studied in the past three decades, and it is found that the propensity of STRs to adopt non-B DNA structures depends their sequence contexts. For instances, the CAG/CTG repeats formed hairpins [23], the guanine-rich CGG and GGGGCC repeats formed G-quadruplexes [24,25], and the cytosine-rich CCG and CCCCGG formed i-motifs [26] (Fig. 1).

Expansions of pyrimidine-rich CCTG, ATTCT and ATTTT/ATTTC repeats are associated with myotonic dystrophy type 2 (DM2) [27], spinocerebellar ataxia type 10 (SCA10) [28], and SCA37 and six subtypes of familial adult myoclonic epilepsy (FAME) [29], respectively. An earlier chemical/enzymatic probing study showed the hairpin formation propensity of CCTG repeats [30], and a subsequent NMR study revealed that CCTG repeats formed hairpin and dumbbell structures [31]. ATTCT repeats have been reported to form unwound secondary structures which may cause repeat expansions via replication re-initiations [32]. However, knowledge of these pyrimidine-rich STRs rests on their secondary structures, with a lacking of high-resolution structural information. Recently, the discovery of a new type of non-B DNA structure, namely the minidumbbell (MDB), provides new structural insights into these pyrimidine-rich sequences. In 2015, Guo and Lam first reported the MDB structures formed in TTTA and CCTG repeats [33,34], and recently they found a similar type of MDB structure formed in ATTCT repeats [35]. These findings help to decipher mechanisms of repeat expansions in these pyrimidine-rich STRs. In this review, we first introduce the MDBs formed in natural STRs with a focus on their high-resolution structures revealed by solution nuclear magnetic resonance (NMR) spectroscopy, then discuss the effects of sequence context, chemical environment, and modified nucleobase on structure and thermostability of MDBs, and finally present perspectives on exploring the sequence criteria and biological functions of MDBs.
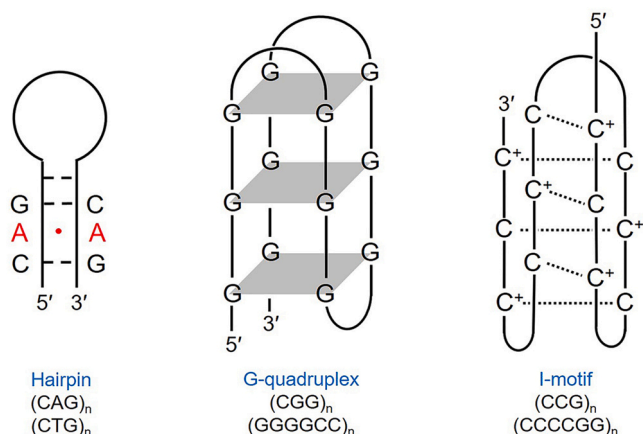
2. MDB structures formed in natural pyrimidine-rich repeats

2.1. Type I and type II tetraloops

There are two well-defined folding geometries of DNA tetraloops, namely, the type I and type II loops (Fig. 2 A) [36]. To facilitate discussions, here we named the first, second, third and fourth loop residues as L1, L2, L3 and L4, respectively. In a type I loop, L1 and L4 form a loop-closing base pair, and L1 to L3 continuously stack at the 5′-end of the loop. In a type II loop, L1 and L4 form a loop-closing base pair, L2 folds into the minor groove, and L3 stacks on L1-L4. Type I and type II loops are found in non-B DNAs such as hairpins and three-way junctions [37–41]. With reference to the NMR features of B-DNA, both type I and type II loops exhibit unusually upfield proton (^{1}H) and phosphorous (^{31}P) NMR signals of L3 [36]. For the type II loop, the ^{1}H and ^{31}P NMR signals of L2 are unusually downfield shifted [36].

2.2. MDB structures in TTTA and CCTG tetranucleotide repeats

Two MDB-like structures formed by cyclic octamers d(pCATTCATT) and d(pTGCTCGCT) were reported by Escaja et al. in 2003 [42]. In each octamer, the 5′ and 3′ terminal residues were covalently linked via a cyclization reaction. Solution NMR structure of d(pCATTCATT) revealed an MDB-like conformer composed of two TTCA type II tetraloops. In 2015, Guo and Lam discovered the MDB structure formed in natural TTTA tetranucleotide repeats [33] which were located in the coding region of the intercellular adhesin C (*icaC*) gene of *Staphylococcus aureus*. This MDB contained two TTTA repeats, which folded into two type II tetraloops, respectively. Herein, this structure was called as the TTTA MDB. The solution NMR structure of this MDB showed that T1-A4 and T5-A8 formed Watson-Crick loop-closing base pairs, T2 and T6 folded into the minor groove, whereas T3 and T7 stacked on T1-A4 and T5-A8, respectively (Fig. 2B, left) [43]. The TTTA MDB was highly compact and exhibited multiple loop-loop interactions, including base-base stackings between T1-A4 and T5-A8, base-base stacking and hydrogen bonds between T2 and T6, and hydrogen bonds between T2 and T5/T7. Besides, there were hydrophobic interactions between T3 and T1/T2, and T7 and T5/T6 (Fig. 2B, right).

CCTG tetranucleotide repeats are located in intron 1 of the human zinc finger protein 9 (*ZNF9*) gene. Expansion from a normal range of ~11–30 repeats to a pathogenic range of ~75–11000 repeats is associated with DM2 [27]. A subsequent solution NMR study demonstrated the formation of MDB structure in CCTG repeats [34]. The CCTG MDB was composed of two repeats, which folded into two type II tetraloops, respectively. In this MDB, C1-G4 and C5-G8 formed two Watson-Crick loop-closing base pairs, C2 and C6 folded into the minor groove, whereas the T3 and T7 stacked on the loop-closing base pairs (Fig. 2 C, left). The overall structure of the CCTG MDB was similar to that of the TTTA MDB, but it had the two minor groove residues forming a C·C mismatch instead of stacking with each other [43]. The CCTG MDB also exhibited extensive stabilizing interactions among loop residues, including base-base stackings, hydrogen bonds and hydrophobic interactions (Fig. 2 C, right).

The energy landscapes of the TTTA and CCTG MDBs were further investigated by Wales's group using molecular dynamic (MD) simulations [44]. Their results showed that the CCTG MDB was the most stable structure while the TTTA MDB was not the lowest-energy structure. Some TTTA MDB-like structures with T1-A4 or T5-A8 Hoogsteen base pair were found to have the lowest energies. Therefore, they suggested that the energy landscapes of the TTTA and CCTG MDBs were mainly determined by the stability of L1-L4 and L1′-L4′ base pairs. This was consistent with the result of an experimental study on how sequence context affected MDB thermostability as discussed in Section 3.1.

Based on the discovery of MDB structures, a 3′-strand slippage model has been proposed to explain TTTA and CCTG repeat

Fig. 1. Schematics of the hairpin, G-quadruplex and i-motif structures formed by STRs.
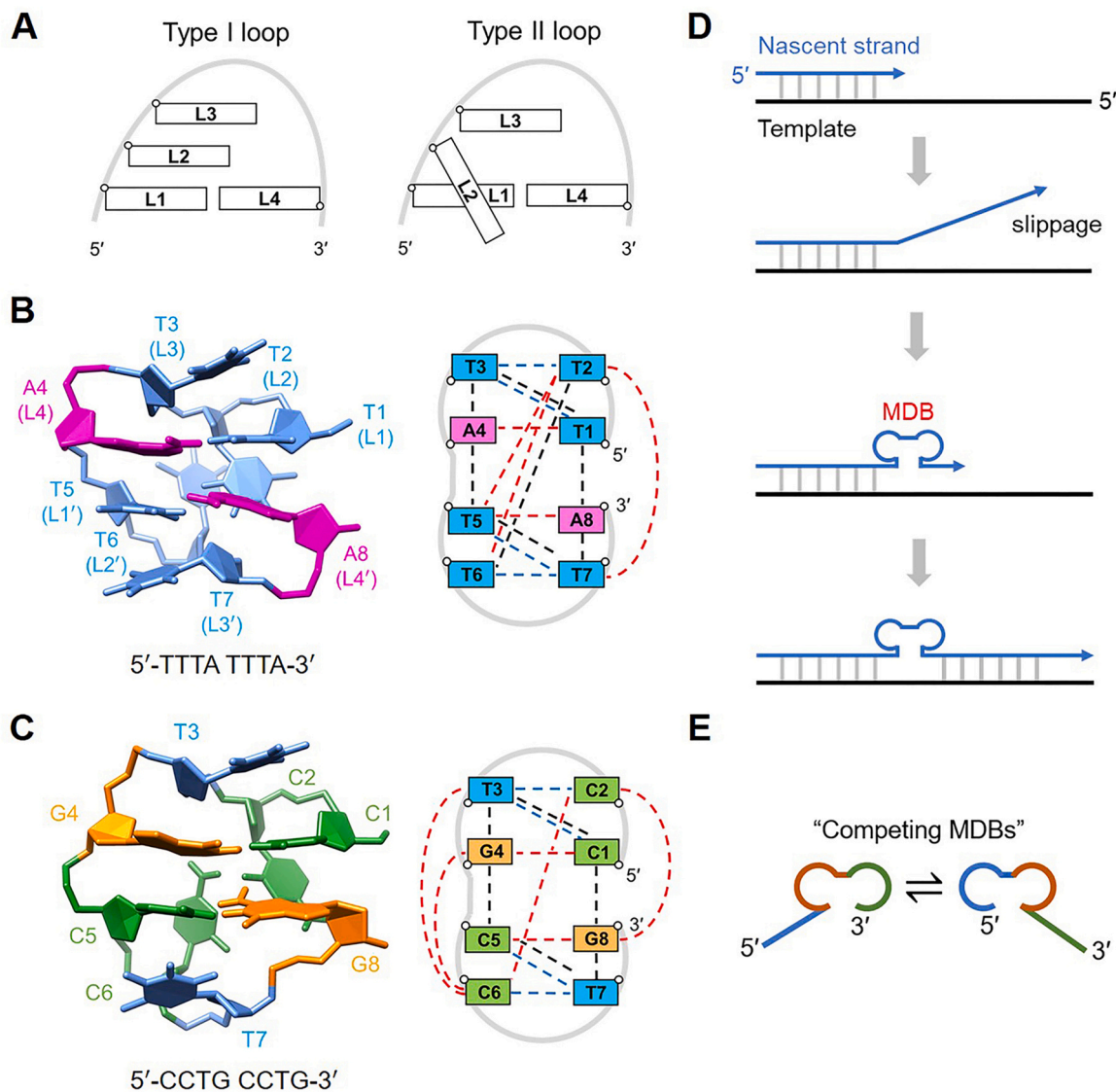
Fig. 2. (A) Schematics of type I and type II tetraloops. (B) Solution NMR structure and schematic of the TTTA MDB (PDB ID: 5GWQ). (C) Solution NMR structure and schematic of the CCTG MDB (PDB ID: 5GWL). Base-base stacking, hydrogen bond and hydrophobic interactions are represented by black, red and blue dashed lines, respectively. (D) The 3′-strand slippage model shows that an MDB forming in the nascent strand during DNA replication can cause a two-repeat expansion. (E) In a segment of three TTTA or CCTG repeats, two adjacent repeats can form competing MDBs which undergo conformational exchange.

expansions in the *icaC* gene of *Staphylococcus aureus* variants and the *ZNF9* gene of DM2 patients, respectively (Fig. 2D) [33,34,45]. During DNA replication, an MDB structure can form in the nascent strand upon a strand slippage at the 3′-end. If the MDB escapes from the proofreading by high-fidelity DNA polymerase during replication, and further escapes from mismatch repair (MMR) proteins during post-replication repair, it will be embedded into the nascent strand and cause a small-scale expansion of two repeats. Multiple small-scale expansions may eventually lead to a large-scale expansion [46].

2.3. Conformational dynamics of MDBs in long tracts of TTTA and CCTG repeats

High-fidelity DNA polymerase and MMR proteins recognize replication errors, such as the mismatch and non-B DNA, using their specific DNA binding sites for the mismatched or unpaired nucleotides [47,48]. Conformational dynamics of non-B DNA may interfere with the recognition by DNA polymerase and MMR proteins, thus affecting the proofreading and repair efficiency, respectively [49,50]. Therefore, elucidating the conformational dynamics of non-B DNA in STRs helps

to improve our understanding on the molecular mechanisms of repeat expansions and facilitates to develop therapeutic strategies [51].

For a long tract of TTTA repeats, Guo et al. reported an interesting phenomenon that two TTTA MDB structures were preferentially formed in the last three repeats at the 3′-end [45]. The two MDBs underwent a conformational exchange and thus they were called as competing MDBs (Fig. 2E). The competing MDBs were also demonstrated to form in a long tract of CCTG repeats [34]. Occurrence of competing MDBs is possible because of the relatively low thermostability of TTTA and CCTG MDBs (melting temperature T_m of ∼20 °C) [33,34]. Such conformational dynamics may facilitate the MDB to escape from being repaired by DNA polymerase and MMR proteins.

2.4. MDB structure in ATTCT pentanucleotide repeats

ATTCT pentanucleotide repeat expansions in intron 9 of the *ATXN10* gene are associated with SCA10. The numbers of repeats in normal individuals and SCA10 patients are ∼10–22 and ∼850–4500, respectively [28]. For a long time, ATTCT repeats had been considered as DNA unwinding elements which could not form stable
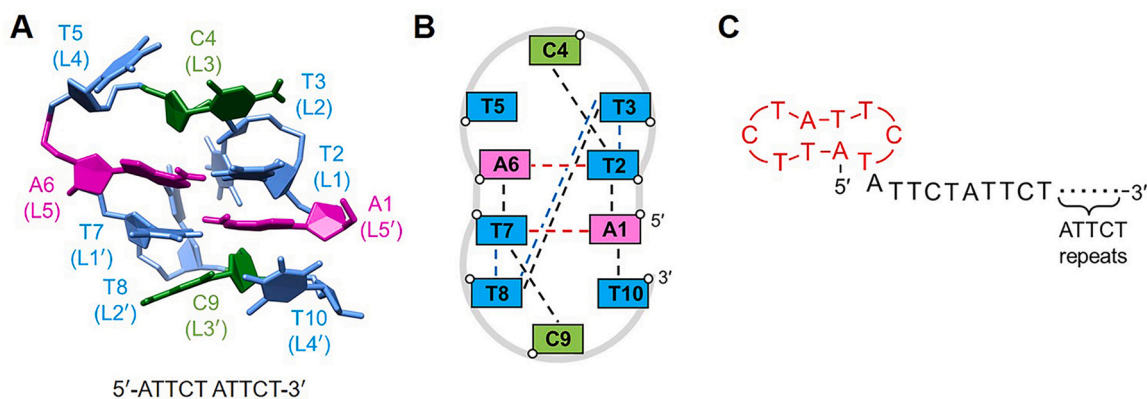
Fig. 3. (A) Solution NMR structure and (B) schematic of ATTCT MDB (PDB ID: 6IY5). In the schematic, base-base stacking, hydrogen bond and hydrophobic interactions are represented by black, red and blue dashed lines, respectively. (C) The MDB was preferentially formed at the 5′-end in a long tract of ATTCT repeats.

secondary structures, and thus ATTCT repeat expansions were explained by replication re-initiations [52] and template switching models [53]. In a recent study, Guo et al. determined the solution NMR structure of two ATTCT repeats which folded into an MDB [35]. This MDB was composed of a TTCTA regular pentaloop and a TTCT/A quasi pentaloop. As defined by Wijmenga et al., the quasi loop has a similar folding geometry as a regular loop but with a backbone discontinuous site (represented by a symbol of "/") between two adjacent loop residues [39]. In the regular loop formed by T2, T3, C4, T5 and A6, T2-A6 adopted a Watson-Crick base pair, T3 folded into the minor groove, C4 stacked on T2-A6, and T5 was located in the major groove. In the quasi loop formed by T7, T8, C9, T10 and A1, T7-A1 adopted a Watson-Crick base pair, T8 folded into the minor groove, C9 stacked on T7-A1, and T10 stacked well with A1 (Fig. 3 A). The stabilizing forces, including base-base stackings, hydrogen bonds and hydrophobic interactions, in the ATTCT MDB are shown in Fig. 3B.

Intriguingly, the MDB was preferentially formed at the 5′-end in a long tract of ATTCT repeats (Fig. 3 C) [35]. The MDB could also be sustained in the presence of flanking sequences, such as a 5′-T, -CT, -TCT, -TTCT or a 3′-A, -AT, -ATT and -ATTC. The 3′-flanking sequences showed less destabilizing effect on the MDB than 5′-flanking sequences, which explains the preferential formation of an MDB at the 5′-termini in longer ATTCT repeats. The capability of ATTCT MDB escaping from DNA polymerase's proofreading was investigated by in vitro primer extension assay [35], revealing that the MDB formed at 5-bp or further away from the priming site could escape from the proofreading by DNA polymerase, and eventually retained in the primer and caused a two-repeat expansion. When the MDB was formed at 3- or 4-bp away from the priming sites, it could stall the polymerase for primer extensions. The discovery of ATTCT MDB provides an alternative pathway for repeat expansions in SCA10 apart from the replication re-initiation and template switching models [52,53]. Very recently, Li et al., have reported that the ATTTT/ATTTC repeats associated with SCA37 and FAMEs formed MDB structures [54] that are similar to the ATTCT MDB, further raising the potential biological significance of MDB structures.

2.5. Biological implications of MDB structures in repeat expansions

The model shown in Fig. 2D elucidates a possible pathway for a small-scale repeat expansion via forming MDB during DNA replication. Multiple small-scale expansions may eventually lead to a large-scale expansion [46]. There are also other possible pathways such as replication re-initiations and template switching to explain large-scale expansions during DNA replication [52,53]. In addition, repeat expansions in post-mitotic neurons are suggested to occur via MMR-

dependent pathways, such as the base excision repair [55]. It will be interesting to explore the roles of MDB in repeat expansions in non-dividing cells.

As the formation of an MDB simply requires two repeats, an interesting question is raised that why short repeats in normal individuals do not expand? For autosomal dominant repeat expansion diseases, the normal allele usually contains short repeats. Once the number of repeats reaches a certain threshold, repeats become unstable and prone to expand. For instance, the threshold numbers of CCTG repeats in the *ZNF9* gene associated with DM2 and ATTCT repeats in the *ATXN10* gene associated with SCA10 are around 75 and 750 repeats, respectively [27,28]. Although exact mechanisms of the threshold-dependent genetic instability remain elusive so far, Lee and McMurray have proposed a DNA-mediated model [56]. Upon a strand slippage during DNA replication, the formation of complementary duplex by slipped nascent strand and template, and the formation of non-B DNA by slipped nascent strand will compete. If the rate of forming complementary duplex is slower than the rate of folding a non-B DNA, the non-B DNA will be retained in nascent strand and subsequently lead to repeat expansions. When the number of repeats exceeds a threshold, the rate of forming complementary duplex become slow, raising the frequency of repeat expansions. The MDB structure is small and its folding is expected to be much faster than complementary duplex in longer repeats, so that to cause repeat expansions especially when the number of repeats exceeds a threshold.

3. Effects of sequence context on structure and thermostability of MDBs

Structure and thermostability of non-B DNA are known to depend on sequence context. Based on the originally identified MDB-forming sequences, i.e. TTTA, CCTG, and ATTCT repeats, a series of studies have been conducted to systematically reveal the effects of sequence context on structure and thermostability of MDBs containing (i) two regular tetraloops [57–59], (ii) two regular pentaloops [60], (iii) a regular and a quasi tetraloops, and (iv) a regular and a quasi pentaloops [60,61].

3.1. Effects of sequence context on MDBs containing two regular tetraloops

The expandable CCTG repeats and the TTTA repeats share the same pyrimidine-rich sequence context YYYR (Y and R indicate pyrimidine and purine, respectively). This inspired a comprehensive study on 64 sequences of 5′-CYYG CYYG-3′, 5′-CYYG TYYA-3′, 5′-TYYA CYYG-3′ and 5′-TYYA TYYA-3′ (Y = C or T) which would potentially form MDBs containing two regular type II tetraloops
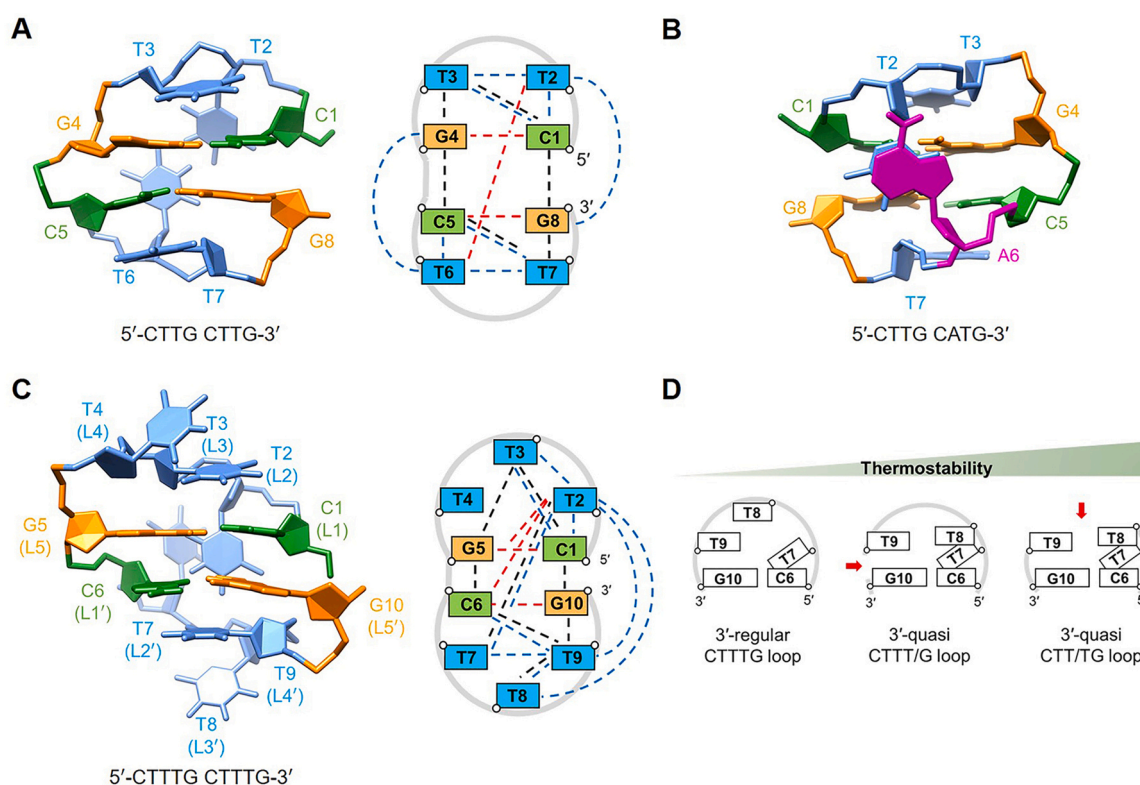
Fig. 4. (A) Solution NMR structure and schematic of the CTTG MDB (PDB ID: 6J37). (B) Solution NMR structure of 5′-CTTG CATG-3′ MDB (PDB ID: 6M0C). (C) Solution NMR structure and schematic of CTTTG MDB (PDB ID: 7VM9). In the schematics of MDBs, base-base stacking, hydrogen bond and hydrophobic interactions are represented by black, red and blue dashed lines, respectively. (D) Schematics of 3′-regular and quasi pentaloops. A 3′-quasi pentaloop is more stabilizing than a 3′-regular pentaloop. The backbone discontinuous sites are highlighted by red arrows. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

[58,59]. The results have revealed effects of sequence context on the MDB thermostability at neutral pH: (i) a L1-L4 or L1′-L4′ C-G was more stabilizing than a T-A base pair, (ii) a L2 or L2′ thymine was more stabilizing than a cytosine residue, and (iii) a L3 or L3′ thymine was more stabilizing than a cytosine residue. As a result, the 5′-CTTG CTTG-3′ sequence formed the most stable MDB at neutral pH (T_m of ∼35 °C). Comparing to the CCTG MDB, the higher thermostability of the CTTG MDB was attributed to a T2·T6 base pair with two hydrogen bonds, and hydrophobic interactions involving the methyl groups of T2 and T6 (Fig. 4 A) [62].

In hairpins and three-way junctions, type II loops were all found to be YYYR and YYYY sequences [37–39], while type I loops were exclusively favored by YRYR and YRRR sequences [40,41]. This is because a pyrimidine L2 of smaller size can be favorably accommodated in the minor groove, whereas a purine L2 of larger size prefers base-base stackings with L1 and L3. Ngai et al. for the first time demonstrated that a type II loop could be formed by YRYR sequence in the context of MDB structure [57]. They systematically investigated eight sequences of 5′-CTTG CRTG-3′, 5′-CRTG CTTG-3′ and 5′-CRTG CRTG-3′ (R = A or G), and revealed that 5′-CGTG CTTG-3′, 5′-CTTG CGTG-3′, and 5′-CTTG CATG-3′ formed MDBs with T_m of ∼25, ∼24, and ∼32 °C, respectively. All these three MDBs contained a CTTG and a CGTG or CATG type II tetraloops. Solution NMR structure of the MDB formed by 5′-CTTG CATG-3′ shows that A6 was driven into the minor groove via base-base stacking with T2 (Fig. 4B). In addition, a L2′ adenine stacked better than a L2′ guanine to bring a higher thermostability.

### 3.2. Effects of sequence context on MDBs containing two regular pentaloops

A recent study showed that an MDB can also be formed with two regular pentaloops, e.g. 5′-CTTTG CTTTG-3′ [60]. The solution NMR

structure of the MDB containing two CTTTG repeats reveals asymmetric loop folding patterns (Fig. 4 C, left). In the 5′-loop, C1-G5 formed a Watson-Crick base pair, T2 folded into the minor groove, T3 stacked on C1-G5, and T4 was located in the major groove. In the 3′-loop, C6-G10 adopted a Watson-Crick base pair, T7 folded into the minor groove, T9 stacked on C6-G10, and T8 stacked on T9. The effect of sequence context on MDBs composed of two regular pentaloops was examined using 5′-CTTTG CTTTG-3′ as a model sequence, revealing that a L2/L2′, L3/L3′ or L4/L4′ thymine was more stabilizing than a cytosine residue [60,63], and this was attributed to the extensive hydrophobic interactions involving the methyl groups of thymine residues (Fig. 4 C, right). In addition, MDBs could sustain a L2′, L3′ or L4′ purine residue (A or G) with T_m of ∼31–36 °C. A L4′ purine was more stabilizing as it stacked better with L1′-L5′ base pair than a L4′ pyrimidine [60].

3.3. Effects of sequence context on MDBs containing a regular and a quasi tetraloops

Quasi loops provided structural flexibility to DNA three-way junctions [64,65]. The effect of a quasi tetraloop on MDBs was systematically examined by creating different backbone discontinuous sites on the CTTG MDB [61]. NMR results showed that backbone discontinuous site between L3′ and L4′ resulted in an MDB comprising a regular CTTG and a quasi CTT/G tetraloops, demonstrating the possibility of including a quasi tetraloop into the MDB.

Recently, the effect of a quasi pentaloop on MDBs has also been studied via introducing backbone discontinuous sites on the CTTTG MDB [60]. A backbone discontinuous site between L3′ and L4′, or between L4′ and L5′ allowed to form stable MDBs containing a regular CTTTG loop and a quasi CTT/TG or CTTT/G loop, and the T_m
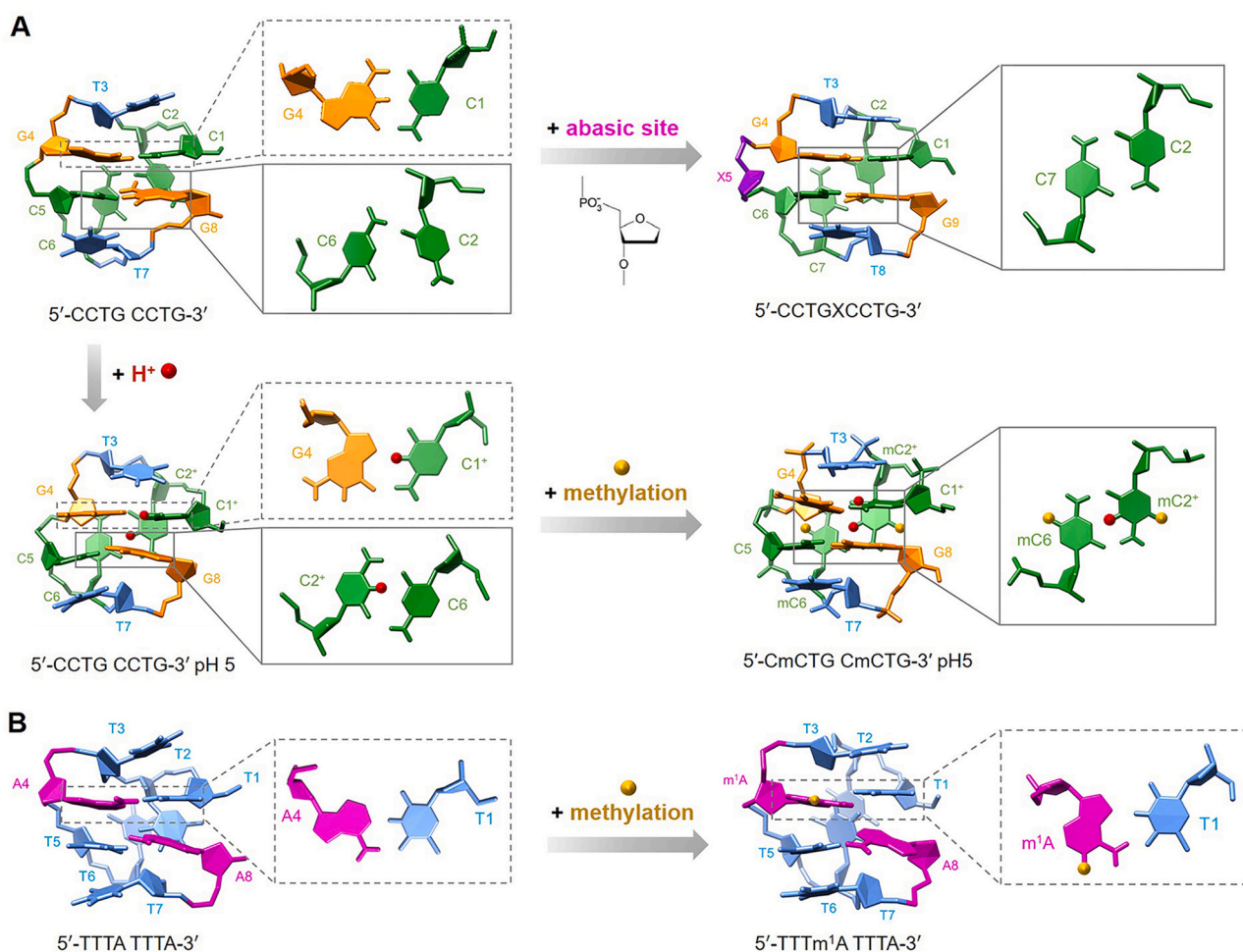
Fig. 5. (A) Solution NMR structures of CCTG MDB at neutral pH (top left), at pH 5 (PDB ID: 7D0Z) (bottom left), with abasic site modification (PDB ID: 6M6K) (top right) and with cytosine methylations (PDB ID: 7D0Y) (bottom right). (B) Solution NMR structures of TTTA MDB without modification (left) and with adenine methylation (PDB ID: 7E4E) (right). Closeup views of the loop-closing base pairs are highlighted by grey-dashed boxes whereas those of the minor groove mispair are highlighted by grey boxes.

values of these two MDBs were ~42 and 34 °C, respectively. These values were higher than that of the CTTTG MDB (T_m of ~32 °C), and this was attributed to a better stacking between L3′ or L4′ with L1′–L5′ base pair in the quasi loop (Fig. 4D).

4. Effects of chemical environment and nucleobase modification on MDBs

4.1. Effect of pH on structure and thermostability of the CCTG MDB

The structure and stability of non-B DNA structures can be affected by pH due to the protonation of adenines and cytosines in an acidic microenvironment [66,67]. For the CCTG MDB, its structure and thermostability changed drastically under different pH conditions [68]. At pH 7, the CCTG MDB contained C1-G4 and C5-G8 Watson-Crick base pairs, and a C2·C6 mispair with predominantly a single hydrogen bond [43]. At pH 5, protonation of C1 and C2 lead to the formation of C1$^+$-G4 Hoogsteen base pair and a three-hydrogen-bond C2$^+$·C6 mispair (Fig. 5A). The terminal residue C1$^+$ showed cation-π interaction with G8, whereas the C2$^+$·C6 mispair contained three hydrogen bonds which undoubtedly provided a substantial stabilization to the MDB. The structural changes of CCTG MDB from pH 7 to pH 5 lead to an increase of T_m by ~24 °C. While the CCTG MDB has been proposed to a structural intermediate that may cause CCTG repeat expansions in DM2 [34], the genetic instability of CCTG

repeats in DM2 patients might be influenced by pH variation in the chemical microenvironment.

4.2. Effect of an abasic site on structure and thermostability of MDBs

Abasic site, which lacks the nucleobase comparing to natural nucleotides, is a frequent DNA lesion occurring at ~10 000 lesions per human cell per day [69]. It appears to be mutagenic by causing base misincorporation, frameshift and repeat expansions [70,71]. The abasic site has been shown to destabilize DNA duplexes and G-quadruplexes [72,73]. When an abasic site was introduced between two tetraloops of an MDB, it did not disturb the overall folding geometry, but significantly enhanced the thermostability of the MDB [74]. Specifically, the T_m of the 5′-CTTGXCTTG-3′ (X represents an abasic site), 5′-CCTGXCCTG-3′ and 5′-TTTAXTTTA-3′ MDBs significantly increased by ~14, 15 and 19 °C than those of the CTTG, CCTG, and TTTA MDBs, respectively. The stabilization was attributed to an increased backbone flexibility upon introducing an abasic site. For instance, as revealed from the solution NMR structure of the 5′-CCTGXCCTG-3′ MDB (Fig. 5A), the backbone flexibility allowed the MDB to form a C·C mispair with an additional hydrogen bond and Na$^+$-mediated electrostatic interactions. Moreover, the abasic site formed additional hydrophobic contacts with neighboring nucleotides using its 2′-methylene groups. In general, abasic site plays a role in the stabilization of MDBs.

4.3. Effect of cytosine methylation on structure and thermostability of MDBs

DNA methylation is an epigenetic mark that regulates gene expression [75]. 5-methylcytosine (mC) is the most abundant DNA methylation in mammalian cells and accounts for ~3% of cytosines in genomic DNA [69]. It has been reported that mC stabilizes a variety of non-B DNA structures including hairpins [76], triplexes [77], i-motifs [78], and G-quadruplexes [79].

Effects of mC at the L1 and L1′ positions was revealed from the 5′-mCTTGXmCTTG-3′ and 5′-CTTGXCTTG-3′ MDBs at pH 7 [80]. The former had a higher T_m by ~11 °C, which was because mC1 and mC5 brought about favorable hydrophobic interactions and more stable loop-closing base pairs. Effect of mC at the L2 and L2′ positions was examined using 5′-CmCTGCmCTG-3′and 5′-CCTGCCTG-3′ at pH 5 [80]. The former had a higher T_m by ~17 °C, which was attributed to the mC2$^+$·mC6 mispair with additional hydrophobic interactions with loop-closing base pairs, as well as favorable electrostatic interactions and improved hydrogen bond geometries (Fig. 5 A). At present, the L1/L1′/L2/L2′-methylated 5′-mCmCTGmCmCTG-3′ MDB at pH 5 remains the most stable MDB ever reported (T_m of ~69 °C). In short, 5mC enhanced the thermostability of MDBs by enabling more stable base pairs, hydrophobic contacts and improved base-base stackings. The stabilization effect might affect the propensity of CCTG repeat expansions in DM2.

4.4. Effect of adenine methylation on structure and thermostability of MDBs

N^6-methyladenine (m^{6}A) is the most prevalent methylated base in prokaryotes with gene regulatory functions, though its abundance in eukaryotes remains a debate [81]. The occurrence of m^{6}A has been reported to destabilize duplexes [82] and G-quadruplexes [83]. N^1-methyladenine (m^{1}A) is another methylated adenine caused by alkylating agents, which can block DNA replication if not repaired [84]. It has been reported that the substitution of an adenine by m^{1}A in the duplex turns the T-A Watson-Crick base pair to T·m^{1}A Hoogsteen base pair and destabilizes the duplex [85].

The effect of adenine methylation on MDBs was investigated by substituting A4 in the 5′-TTTATTTA-3′ MDB with m^{6}A or m^{1}A [86]. The m^{6}A-substituted sequence 5′-TTTm^{6}ATTTA-3′ was found to form two MDB conformers. The major MDB conformer had a T1·m^{6}A4 Hoogsteen base pair while the minor MDB conformer exhibited a T1-m^{6}A4 Watson-Crick base pair. The T_m of the major MDB conformer was ~9 °C lower than that of the unmodified TTTA MDB. For the m^{1}A-substituted sequence 5′-TTTm^{1}ATTTA-3′, it adopted a single MDB conformer with a T1·m^{1}A Hoogsteen base pair (Fig. 5B), and its T_m was ~4 °C higher than that of the unmodified TTTA MDB. This represents the first report of the stabilization effect of m^{1}A on a DNA secondary structure.

5. Summary and outlook

The discovery of MDB structure has provided new insights into the genetic instability of pyrimidine-rich repeats associated neurodegenerative diseases such as DM2 and SCA10. Efforts have also been made to understand the effects of sequence context, nucleobase modification and chemical environment on the structure and thermostability of MDBs. However, our understanding on this relatively new type of non-B DNA is still far from mature. First, the sequence criteria for MDB have not been well established. For instance, as all the reported MDBs have C-G or T-A loop-closing base pairs, it is curious if a mismatch can serve as the loop-closing base pair. Second, mapping the MDB-forming hotspots in the human genome and studying their correlations with biological functions need to be considered in the future work. Finally, small-molecule ligands

designed for probing or visualizing MDB structure in vivo will definitely facilitate us to explore the biological significance of MDB. Although the biological significance of this new type of non-B DNA remains elusive at the current stage, an illustration of its structures and conformational dynamics may provide useful information for therapeutic development in the future.

CRediT authorship contribution statement

Yuan Liu: Writing – original draft, Review & editing, Funding acquisition. **Liqi Wan**: Writing – original draft, Review & editing. **Cheuk Kit Ngai**: Writing – original draft, Review & editing. **Yang Wang**: Writing – review & editing. **Sik Lok Lam**: Conceptualization, Supervision. **Pei Guo**: Conceptualization, Supervision, Resources, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare no competing financial interests.

References

[1] Fan H, Chu JY. A brief review of short tandem repeat mutation. Genom Proteom Bioinforma 2007;5:7–14.

[2] Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol 2003;4:R13.

[3] La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 1991;352:77–9.

[4] Orr HT, Chung MY, Banfi S, Kwiatkowski Jr. TJ, Servadio A, et al. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet 1993;4:221–6.

[5] Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet 2018;19:286–98.

[6] Khristich AN, Mirkin SM. On the wrong DNA track: molecular mechanisms of repeat-mediated genome instability. J Biol Chem 2020;295:4134–70.

[7] Depienne C, Mandel JL. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? Am J Hum Genet 2021;108:764–85.

[8] Kim JC, Mirkin SM. The balancing act of DNA repeat expansions. Curr Opin Genet Dev 2013;23:280–8.

[9] Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 1991;65:905–14.

[10] Kremer B, Goldberg P, Andrew SE, Theilmann J, Telenius H, et al. A worldwide study of the Huntington's disease mutation. The sensitivity and specificity of measuring CAG repeats. N Engl J Med 1994;330:1401–6.

[11] Renton AE, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, et al. A hexanucleotide repeat expansion in *C9ORF72* is the cause of chromosome 9p21-linked ALS-FTD. Neuron 2011;72:257–68.

[12] Vital M, Bidegain E, Raggio V, Esperon P. Molecular characterization of genes modifying the age at onset in Huntington's disease in Uruguayan patients. Int J Neurosci 2016;126:510–3.

[13] Morales F, Couto JM, Higham CF, Hogg G, Cuenca P, et al. Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. Hum Mol Genet 2012;21:3558–67.

[14] Florian RT, Kraft F, Leitão E, Kaya S, Klebe S, et al. Unstable TTTTA/TTTCA expansions in MARCH6 are associated with Familial Adult Myoclonic Epilepsy type 3. Nat Commun 2019;10:4919.

[15] Fujino Y, Nagai Y. The molecular pathogenesis of repeat expansion diseases. Biochem Soc Trans 2022;50:119–34.

[16] Stoyas CA, La Spada AR. The CAG-polyglutamine repeat diseases: a clinical, molecular, genetic, and pathophysiologic nosology. Handb Clin Neurol 2018;147:143–70.

[17] Swinnen B, Robberecht W, Van Den Bosch L. RNA toxicity in non-coding repeat expansion disorders. Embo J 2020;39:e101112.

[18] Wojciechowska M, Krzyzosiak WJ. Cellular toxicity of expanded RNA repeats: focus on RNA foci. Hum Mol Genet 2011;20:3811–21.

[19] Todd TW, McEachin ZT, Chew J, Burch AR, Jansen-West K, et al. Hexanucleotide repeat expansions in c9FTD/ALS and SCA36 confer selective patterns of neurodegeneration in vivo. Cell Rep 2020;31:107616.

[20] McEachin ZT, Gendron TF, Raj N, García-Murias M, Banerjee A, et al. Chimeric peptide species contribute to divergent dipeptide repeat pathology in c9ALS/FTD and SCA36. Neuron 2020;107:292–305. e296.

[21] Gacy AM, Goellner G, Juranić N, Macura S, McMurray CT. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell 1995;81:533–40.

[22] Mirkin SM. Expandable DNA repeats and human disease. Nature 2007;447:932–40.

[23] Chi LM, Lam SL. Structural roles of CTG repeats in slippage expansion during DNA replication. Nucleic Acids Res 2005;33:1604–17.

[24] Zhou B, Liu C, Geng Y, Zhu G. Topology of a G-quadruplex DNA formed by C9orf72 hexanucleotide repeats associated with ALS and FTD. Sci Rep 2015;5:16673.

[25] Yi J, Wan L, Liu Y, Lam SL, Chan HYE, et al. NMR solution structures of d(GGCCTG) (n) repeats associated with spinocerebellar ataxia type 36. Int J Biol Macromol 2022;201:607–15.

[26] Abou Assi H, Garavís M, González C, Damha MJ. i-Motif DNA: structural features and significance to cell biology. Nucleic Acids Res 2018;46:8038–56.

[27] Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, et al. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. Science 2001;293:864–7.

[28] Matsuura T, Yamagata T, Burgess DL, Rasmussen A, Grewal RP, et al. Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. Nat Genet 2000;26:191–4.

[29] Loureiro JR, Castro AF, Figueiredo AS, Silveira I. Molecular mechanisms in pentanucleotide repeat diseases. Cells 2022;11:205.

[30] Dere R, Napierala M, Ranum LP, Wells RD. Hairpin structure-forming propensity of the (CCTG.CAGG) tetranucleotide repeats contributes to the genetic instability associated with myotonic dystrophy type 2. J Biol Chem 2004;279:41715–26.

[31] Lam SL, Wu F, Yang H, Chi LM. The origin of genetic instability in CCTG repeats. Nucleic Acids Res 2011;39:6260–8.

[32] Liu G, Bissler JJ, Sinden RR, Leffak M. Unstable spinocerebellar ataxia type 10 (ATTCT*(AGAAT) repeats are associated with aberrant replication at the ATX10 locus and replication origin-dependent expansion at an ectopic site in human cells. Mol Cell Biol 2007;27:7828–38.

[33] Guo P, Lam SL. Unusual structures of TTTA repeats in *icaC* gene of *Staphylococcus aureus*. FEBS Lett 2015;589:1296–300.

[34] Guo P, Lam SL. New insights into the genetic instability in CCTG repeats. FEBS Lett 2015;589:3058–63.

[35] Guo P, Lam SL. Minidumbbell structures formed by ATTCT pentanucleotide repeats in spinocerebellar ataxia type 10. Nucleic Acids Res 2020;48:7557–68.

[36] Lam SL, Chi LM. Use of chemical shifts for structural studies of nucleic acids. Prog Nucl Magn Reson Spectrosc 2010;56:289–310.

[37] van Dongen MJ, Wijmenga SS, van der Marel GA, van Boom JH, Hilbers CW. The transition from a neutral-pH double helix to a low-pH triple helix induces a conformational switch in the CCCG tetraloop closing a Watson-Crick stem. J Mol Biol 1996;263:715–29.

[38] Cho SH, Chin KH, Chen CW. Enhanced loop DNA folding induced by thymine-CH3 group contact and perpendicular guanine-thymine interaction. J Biomol NMR 2001;19:33–48.

[39] Wu B, Girard F, van Buuren B, Schleucher J, Tessari M, et al. Global structure of a DNA three-way junction by solution NMR: towards prediction of 3H fold. Nucleic Acids Res 2004;32:3228–39.

[40] Ippel HH, van den Elst H, van der Marel GA, van Boom JH, Altona C. Structural similarities and differences between H1- and H2-family DNA minihairpin loops: NMR studies of octameric minihairpins. Biopolymers 1998;46:375–93.

[41] Avizonis DZ, Kearns DR. Structural characterization of d(CAACCCGTTG) and d (CAACGGGTTG) mini-hairpin loops by heteronuclear NMR: the effects of purines versus pyrimidines in DNA hairpins. Nucleic Acids Res 1995;23:1260–8.

[42] Escaja N, Gómez-Pinto I, Rico M, Pedroso E, González C. Structures and stabilities of small DNA dumbbells with Watson-Crick and Hoogsteen base pairs. Chembiochem 2003;4:623–32.

[43] Guo P, Lam SL. Minidumbbell: a new form of native DNA structure. J Am Chem Soc 2016;138:12534–40.

[44] Klimavicz JS, Röder K, Wales DJ. Energy landscapes of mini-dumbbell DNA octanucleotides. J Chem Theory Comput 2018;14:3870–6.

[45] Guo P, Lam SL. The competing mini-dumbbell mechanism: new insights into CCTG repeat expansion. Signal Transduct Target Ther 2016;1:16028.

[46] Guo P, Lam SL. Unusual structures of CCTG repeats and their participation in repeat expansion. Biomol Concepts 2016;7:331–40.

[47] Reha-Krantz LJ. DNA polymerase proofreading: multiple roles maintain genome stability. Biochim Biophys Acta 2010;1804:1049–63.

[48] Natrajan G, Lamers MH, Enzlin JH, Winterwerp HH, Perrakis A, et al. Structures of Escherichia coli DNA mismatch repair enzyme MutS in complex with different mismatches: a common recognition mode for diverse substrates. Nucleic Acids Res 2003;31:4814–21.

[49] Panigrahi GB, Slean MM, Simard JP, Gileadi O, Pearson CE. Isolated short CTG/CAG DNA slip-outs are repaired efficiently by hMutSbeta, but clustered slip-outs are poorly repaired. Proc Natl Acad Sci USA 2010;107:12593–8.

[50] Slean MM, Reddy K, Wu B, Nichol Edamura K, Kekis M, et al. Interconverting conformations of slipped-DNA junctions formed by trinucleotide repeats affect repair outcome. Biochemistry 2013;52:773–85.

[51] Hu T, Morten MJ, Magennis SW. Conformational and migrational dynamics of slipped-strand DNA three-way junctions containing trinucleotide repeats. Nat Commun 2021;12:204.

[52] Potaman VN, Bissler JJ, Hashem VI, Oussatcheva EA, Lu L, et al. Unpaired structures in SCA10 (ATTCT)n.(AGAAT)n repeats. J Mol Biol 2003;326:1095–111.

[53] Cherng N, Shishkin AA, Schlager LI, Tuck RH, Sloan L, et al. Expansions, contractions, and fragility of the spinocerebellar ataxia type 10 pentanucleotide repeat in yeast. Proc Natl Acad Sci USA 2011;108:2843–8.

[54] Li J, Wan L, Wang Y, Chen Y, Lee HK, et al. Solution nuclear magnetic resonance structures of ATTTT and ATTTC pentanucleotide repeats associated with SCA37 and FAMEs. ACS Chem Neurosci 2023;14:289–99.

[55] Gonitel R, Moffitt H, Sathasivam K, Woodman B, Detloff PJ, et al. DNA instability in postmitotic neurons. Proc Natl Acad Sci USA 2008;105:3467–72.

[56] Lee DY, McMurray CT. Trinucleotide expansion in disease: why is there a length threshold? Curr Opin Genet Dev 2014;26:131–40.

[57] Ngai CK, Lam SL, Lee HK, Guo P. High-resolution structures of DNA minidumbbells comprising type II tetraloops with a purine minor groove residue. J Phys Chem B 2020;124:5131–8.

[58] Liu Y, Lam SL. Sequence effect on the formation of DNA minidumbbells. J Phys Chem B 2017;121:10338–43.

[59] Li M. DNA minidumbbells formed by hybrid type II loops (Master's Thesis). The Chinese University of Hong Kong; 2018. (Master's Thesis). ⟨https://repository.lib.cuhk.edu.hk/tc/item/cuhk-2327303⟩.

[60] Ngai CK, Lam SL, Lee HK, Guo P. A purine and a backbone discontinuous site alter the structure and thermal stability of DNA minidumbbells containing two pentaloops. FEBS Lett 2022;596:826–40.

[61] Liu Y, Guo P, Lam SL. Formation of a DNA mini-dumbbell with a quasi-type II loop. J Phys Chem B 2017;121:2554–60.

[62] Guo P, Lam SL. Unprecedented hydrophobic stabilizations from a reverse wobble T·T mispair in DNA minidumbbell. J Biomol Struct Dyn 2020;38:1946–53.

[63] Ngai CK. New insights into DNA minidumbbell forming sequences (Doctoral Thesis). The Chinese University of Hong Kong; 2021. (Doctoral Thesis). ⟨https://repository.lib.cuhk.edu.hk/tc/item/cuhk-2876339⟩.

[64] van Buuren BNM, Overmars FJJ, Ippel JH, Altona C, Wijmenga SS. Solution structure of a DNA three-way junction containing two unpaired thymidine bases. identification of sequence features that decide conformer selection1 1Edited by I. Timoco. J Mol Biol 2000;304:371–83.

[65] Overmars FJ, Pikkemaat JA, van den Elst H, van Boom JH, Altona C. NMR studies of DNA three-way junctions containing two unpaired thymidine bases: the influence of the sequence at the junction on the stability of the stacking conformers. J Mol Biol 1996;255:702–13.

[66] Zheng LL, Li JZ, Li YX, Gao JB, Dong JX, et al. pH-Responsive DNA motif: from rational design to analytical applications. Front Chem 2021;9:732770.

[67] Nishio M, Tsukakoshi K, Ikebukuro K. G-quadruplex: flexible conformational changes by cations, pH, crowding and its applications to biosensing. Biosens Bioelectron 2021;178:113030.

[68] Guo P, Lam SL. An extraordinarily stable DNA minidumbbell. J Phys Chem Lett 2017;8:3478–81.

[69] Lindahl T. Instability and decay of the primary structure of DNA. Nature 1993;362:709–15.

[70] Blanca G, Villani G, Shevelev I, Ramadan K, Spadari S, et al. Human DNA polymerases lambda and beta show different efficiencies of translesion DNA synthesis past abasic sites and alternative mechanisms for frameshift generation. Biochemistry 2004;43:11605–15.

[71] Lyons-Darden T, Topal MD. Abasic sites induce triplet-repeat expansion during DNA replication in vitro. J Biol Chem 1999;274:25975–8.

[72] Minetti CA, Sun JY, Jacobs DP, Kang I, Remeta DP, et al. Impact of bistrand abasic sites and proximate orientation on DNA global structure and duplex energetics. Biopolymers 2018;109:e23098.

[73] Esposito V, Martino L, Citarella G, Virgilio A, Mayol L, et al. Effects of abasic sites on structural, thermodynamic and kinetic properties of quadruplex structures. Nucleic Acids Res 2010;38:2069–80.

[74] Wan L, Lam SL, Lee HK, Guo P. Rational design of a reversible Mg(2+)/EDTA-controlled molecular switch based on a DNA minidumbbell. Chem Commun (Camb) 2020;56:10127–30.

[75] Breiling A, Lyko F. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. Epigenetics Chromatin 2015;8:24.

[76] Xodo LE, Manzini G, Quadrifoglio F, van der Marel G, van Boom J. DNA hairpin loops in solution. Correlation between primary structure, thermostability and reactivity with single-strand-specific nuclease from mung bean. Nucleic Acids Res 1991;19:1505–11.

[77] Leitner D, Schröder W, Weisz K. Influence of sequence-dependent cytosine protonation and methylation on DNA triplex stability. Biochemistry 2000;39:5886–92.

[78] Wright EP, Abdelhamid MAS, Ehiabor MO, Grigg MC, Irving K, et al. Epigenetic modification of cytosines fine tunes the stability of i-motif DNA. Nucleic Acids Res 2020;48:55–62.

[79] Lin J, Hou JQ, Xiang HD, Yan YY, Gu YC, et al. Stabilization of G-quadruplex DNA by C-5-methyl-cytosine in bcl-2 promoter: implications for epigenetic regulation. Biochem Biophys Res Commun 2013;433:368–73.

[80] Wan L, Yi J, Lam SL, Lee HK, Guo P. 5-Methylcytosine substantially enhances the thermal stability of DNA minidumbbells. Chemistry 2021;27:6740–7.

[81] Xiao CL, Zhu S, He M, Chen D, Zhang Q, et al. N(6)-Methyladenine DNA modification in the human genome. Mol Cell 2018;71:306–18. e307.

[82] Engel JD, von Hippel PH. Effects of methylation on the stability of nucleic acid conformations: studies at the monomer level. Biochemistry 1974;13:4143–58.

[83] Iwasaki Y, Ookuro Y, Iida K, Nagasawa K, Yoshida W. Destabilization of DNA and RNA G-quadruplex structures formed by GGA repeat due to N(6)-methyladenine modification. Biochem Biophys Res Commun 2022;597:134–9.

[84] Delaney JC, Essigmann JM. Mutagenesis, genotoxicity, and repair of 1-methyladenine, 3-alkylcytosines, 1-methylguanine, and 3-methylthymine in alkB Escherichia coli. Proc Natl Acad Sci USA 2004;101:14051–6.

[85] Yang H, Lam SL. Effect of 1-methyladenine on thermodynamic stabilities of double-helical DNA structures. FEBS Lett 2009;583:1548–53.

[86] Wan L, Lam SL, Lee HK, Guo P. Effects of adenine methylation on the structure and thermodynamic stability of a DNA minidumbbell. Int J Mol Sci 2021;22:3363.