

**A role for non-B DNA forming sequences in mediating micro-lesions
causing human inherited disease**

Mihir Anant Kamat¹, Albino Bacolla², David N Cooper² and Nadia Chuzhanova¹

¹School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham
NG11 8NS, UK.

²Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff
CF14 4XN, UK.

* Corresponding author: Nadia Chuzhanova, School of Science and Technology, Nottingham
Trent University, Clifton Lane, Nottingham, NG11 8NS, UK. Tel: +44 (0)115 848 8304
E-mail: nadia.chuzhanova@ntu.ac.uk

Abstract

Missense/nonsense mutations and micro-deletions/micro-insertions of <21bp together represent ~76% of all mutations causing human inherited disease. Previous studies have shown that their occurrence is influenced by sequences capable of non-B DNA formation (direct, inverted and mirror repeats; G-quartets). We found that a greater than expected proportion (~21%) of both micro-deletions and micro-insertions occur within direct repeats and are explicable by slipped misalignment. A novel mutational mechanism, non-B DNA triplex formation followed by DNA repair, is proposed to explain ~5% of micro-deletions and micro-insertions at mirror repeats. Further, G-quadruplex-forming sequences, direct and inverted repeats appear to play a prominent role in mediating missense mutations, whereas only direct and inverted repeats mediate nonsense mutations. We suggest a mutational mechanism involving slipped strand mispairing, slipped structure formation and DNA repair, to explain ~15% of missense and ~12% of nonsense mutations leading to the formation of perfect direct repeats from imperfect repeats, or to the extension of existing direct repeats. Similar proportions of missense and nonsense mutations were explicable by the mechanism of hairpin loop formation and DNA repair leading to the formation of perfect inverted repeats from imperfect repeats. The proposed mechanisms provide new insights into mutagenesis underlying pathogenic micro-lesions.

Keywords: non-B DNA, missense mutations, nonsense mutations, micro-deletions, micro-insertions, mechanisms of mutagenesis

Introduction

Micro-lesions comprising missense and nonsense mutations, and micro-deletions and micro-insertions of <21bp represent ~76% of all lesions known to cause human inherited disease (Stenson *et al.*, 2014). Previous studies have shown that the occurrence of germline mutations is influenced by the local DNA sequence environment, including the presence of non-B DNA-forming repeats (reviewed in Cooper *et al.*, 2011). It is known that direct repeats, inverted repeats, and mirror repeats comprising runs of purine:pyrimidine bases are capable of folding into non-canonical (non-B) DNA conformations, i.e. slipped, hairpin or cruciform, and triplex structures, respectively (Sinden, 1994), whereas four runs of two or more guanines with an interspaced loop of length 1 to 7bp are capable of G-quadruplex (also known as G4 or G4 DNA) structure formation (Rouleau *et al.*, 2015) (Figure 1).

A combination of *in silico* and *in vivo* studies have revealed that, independent of genomic location, non-B DNA-forming sequences display more genetic variation than their flanking counterparts (Bacolla *et al.*, 2011; Du *et al.*, 2014), implying that such sequences are more mutable than the genome-wide average. Analyses of mutations causing human genetic disease have provided evidence for the occurrence of mutations within non-B DNA-forming sequences (Wells, 2007; Chuzhanova *et al.*, 2009; Bacolla *et al.*, 2011). However, the extent to which they are involved in mediating mutations on a genome-wide scale has not yet been ascertained. Indeed, the association of non-B DNA-forming sequences with genomic instability has been best established in the areas of triplet repeat expansion diseases (Zhao and Usdin, 2015; Iyer *et al.*, 2015), and in several gross chromosomal abnormalities, both in the germline (Cooper *et al.*, 2011; Verdin *et al.*, 2013; You *et al.*, 2013; Wu *et al.*, 2014; Javadekar and Raghavan, 2015) and in cancer (De and Michor, 2011; Nambiar *et al.*, 2013; Jeitany *et al.*, 2015; Lu *et al.*, 2015; Williams *et al.*, 2015). Filling this knowledge gap is of particular interest in the field of medical genetics, given the widespread occurrence of non-B DNA-forming repeats in the

human and other mammalian genomes (Du *et al.*, 2014).

Previous studies based on smaller datasets of micro-deletions and micro-insertions than used herein have attempted to relate their occurrence to the presence of non-B DNA-forming repeats in their immediate vicinity (Cooper and Krawczak, 1993; Ball *et al.*, 2005). Cooper and Krawczak (1993) found that the majority of micro-insertions were flanked by repeat sequences, and that all the micro-deletions analysed occurred within (or were flanked by) direct repeats. Likewise, Ball *et al.* (2005) found that 84% of micro-insertions and 81% of micro-deletions occurred in the vicinity of direct, inverted and mirror repeats. A few studies have also reported the occurrence of single base-pair substitutions within non-B DNA-forming sequences (Spurlock *et al.*, 2009; Tappino *et al.*, 2009). In particular, Spurlock *et al.* (2009) suggested that a G>T:E73X [GAA>TAA, c.217] nonsense mutation within an imperfect direct repeat in the *SPRED1* gene could have been mediated by a slipped structure. A similar mechanism was suggested by Tappino *et al.* (2009) to explain the occurrence of nonsense mutations within imperfect direct repeats in the *GNPTAB* gene, causing mucopolipidosis II alpha or beta. However, no meta-analyses have been undertaken to assess the extent to which non-B DNA-forming sequences are involved in mediating single base-pair substitutions underlying human inherited disease, or to explore their underlying mutational mechanisms. Moreover, although the formation of intramolecular G-quadruplex structures *in vivo* has been reported for some time (Sen and Gilbert, 1988), their involvement in mediating micro-lesions causing human inherited disease genome-wide has not been assessed.

Here, we ascertain the relationship between a large dataset of mutations causing human inherited disease and various types of non-B DNA-forming repeat. We also propose some potential mutational mechanisms that could be responsible for generating non-B DNA-mediated missense and nonsense mutations, micro-deletions and micro-insertions.

Materials and Methods

Mutation Data

In December 2010 (when this study commenced) the HGMD Professional Release comprised 83,751 pathological micro-lesions in the coding regions of 2,447 human genes. These included 47,119 missense mutations, 12,362 nonsense mutations, 17,208 micro-deletions and 7,062 micro-insertions in 2,171, 1,360, 1,536 and 1,156 genes, respectively. The curators of the HGMD provided a collection of cDNA sequences and ‘extended cDNA sequences’, comprising the coding exons, along with ± 40 bp of intronic sequence flanking the splice junctions, plus the 5′ and 3′ regions of ± 40 bp flanking the initiation and termination codons, respectively.

Any missing extended cDNA sequences were obtained from human genome assembly GRCh37 (hg19) employing the Ensembl database versions 67 (May 2012) or 69 (October 2012), available at <http://www.ensembl.org/index.html>. The cDNA sequences were used as a reference for single base-pair substitutions, whereas the extended cDNA sequences were used for the remaining lesions. Mutated nucleotides, flanking nucleotide sequences, and genomic positions were verified against the GRCh37 reference human genome assembly. Instances of micro-deletions with both 5′ and 3′ breakpoints (or micro-insertions with the breakpoint) occurring within a non-coding region, were excluded from the analyses.

Control Datasets

Control datasets matching the single base-pair substitution dataset were generated using a set of 2,171 genes. For each gene, missense mutations were generated randomly such that the number of sequence alterations matched the number of known (i.e. HGMD) missense mutations, and this comprised one simulation. Such simulations were then repeated 1,000 times and the results were averaged.

Control datasets for micro-deletion and micro-insertion breakpoints were generated as follows. For a given gene, the first breakpoint was generated randomly within the extended cDNA sequence. The location of the second breakpoint was then selected so as to match the distribution of the number of deleted bases in the HGMD dataset of micro-deletions. Hence, for each gene, the number of artificially generated micro-deletions matched the number of observed micro-deletions. This procedure was repeated for all genes, such that the number of micro-deletion 5' and 3' breakpoints occurring in the exonic regions matched the number of observed breakpoints in the HGMD dataset. The process was also repeated 1,000 times to avoid any biases. Control datasets for micro-insertions were generated in a similar manner, although only one breakpoint was generated in this case.

Non-B DNA-Forming Sequences

Direct, inverted and R•Y-rich (80%) mirror repeats of length ≥ 5 bp and ≤ 20 bp apart, capable of forming non-B DNA conformations (slipped, hairpin or cruciform structures, intramolecular triplexes (Ball *et al.*, 2005), and G-quartets (four runs of guanines with interspaced loops of length 1-7bp, i.e. $G_n N_{1-7} G_n N_{1-7} G_n N_{1-7} G_n$, where $n = 2, 3$ or 4), capable of G-quadruplex formation (Rouleau *et al.*, 2015), were sought in both the reference and mutated cDNA, and extended cDNA sequences. Single-base substitutions or breakpoints occurring within the non-B DNA forming repeats were termed mutation/breakpoints “in-repeat”; otherwise, they were termed mutations/breakpoints “not in-repeat”.

Bioinformatics Analyses

A novel algorithm based on the principles of complexity analysis (Gusev *et al.*, 1999) and having a linear running time, was designed and implemented in JAVA to identify the different types of repeat within both the cDNA and extended cDNA sequences. The significance of the

findings was then assessed using Fisher's Exact test. A Bonferroni correction was employed to allow for multiple testing.

Results and Discussion

Micro-Deletions and Micro-Insertions Within Non-B DNA-Forming Sequences

Approximately 21% of all micro-deletion 5' breakpoints (3,612/17,208), henceforth called first breakpoints, and micro-insertion breakpoints (1,495/7,062) were found to occur within direct repeats; this proportion is significantly higher than would be expected by chance alone (Fisher's Exact test, $p=1.54\times 10^{-8}$, and $p=4.8\times 10^{-5}$, respectively; Table 1) and is consistent with the known propensity of direct repeats to undergo slipped strand misalignment during DNA replication, which then generates mutations (Levinson and Gutman, 1987; Cooper and Krawczak, 1993; Sinden, 1994; Rosche *et al.*, 1995; Bzymek *et al.*, 1999; Lovett, 2004; Ball *et al.*, 2005). Owing to the close proximity of 3' breakpoints to the first breakpoints in most micro-deletions (84% of breakpoints were <5bp apart), similar results were obtained for the dataset of the 3' breakpoints (results not shown).

Further, ~5% (870) of all analysed micro-deletion first breakpoints and 4.5% (318) of micro-insertion breakpoints occurred within R•Y-rich mirror repeats, resulting in significant over-representation as compared to the control dataset (Fisher's Exact test, $p=3.56\times 10^{-41}$ and $p=6.04\times 10^{-11}$, respectively). Of all micro-deletions and micro-insertions that occurred within mirror repeats, 596 (69%) and 246 (77%) breakpoints, respectively, were found within the mirror repeats themselves. In the remaining cases, we noted the presence of either deleted or inserted fragments that partially overlapped with the mirror repeats and their spacers. The distribution of mutations with respect to the length of the overlaps is given in Supplementary Table 1.

Additionally, for the lesions associated with mirror repeats, 9.4% (1593) of micro-deletions and 12% (888) of micro-insertions were found to occur within 10bp of the repeat boundaries; these proportions are significantly higher than would be expected by chance alone (Fisher's Exact test, $p=4.04\times 10^{-60}$ and $p=9.02\times 10^{-108}$, respectively). We conclude that mirror repeats promote mutagenesis, both within the repeats and in their immediate vicinity.

Our study found that 56% (3,988) and 54% (9,191) of micro-insertions and micro-deletions, respectively, occurred either within or in the immediate vicinity (± 10 bp) of either mirror repeats or direct repeats. These proportions are significantly higher than expected by chance alone ($p < 10^{-149}$), but are smaller than previously reported, viz. 63% for micro-deletions and 69% for micro-insertions, in the vicinity of mirror repeats (Ball *et al.*, 2005). This difference is likely to be due solely to our employing a more stringent definition, which limited mirror repeats to those that are also R•Y-rich ($\geq 80\%$).

As mentioned above, although mirror repeats have been previously implicated in the generation of micro-deletions and micro-insertions, the underlying mechanism has remained unclear (Sinden and Wells, 1992). Cooper and Krawczak (1991, 1993) proposed a mechanism based on the formation of a secondary structure intermediate, i.e. a Moebius loop, although the ability of a Moebius loop to form *in vivo* has not been proven. Conversely, R•Y-rich mirror repeats of the type analysed here are known to fold into intramolecular triplex structures. Hence, to account for the high frequencies of micro-deletions and micro-insertions, both within and flanking the R•Y-rich mirror repeats, we propose a novel mutational mechanism based upon the formation of triplex DNA (H-DNA) coupled with DNA repair, as described below.

A Model for Triplex-Induced Mutagenesis Promoting Micro-Deletion and Micro-Insertion

Although several pathways may be operative in the context of triplex DNA-induced mutation (Belotserkovskii *et al.*, 2007; Wang and Vasquez, 2014), we favor one that is

dependent upon DNA replication (Figure 2). Triplex DNA has been shown to impede progression of an incoming replication complex, thereby leading to double-strand breaks, in a manner that depends on both the length of the structure and the amount of negative superhelical tension present in the DNA template (Patel *et al.*, 2004). Previous work has shown that negative supercoiling is required for all B- to non-B DNA transitions, such that higher levels of negative supercoiling are expected to increase the frequency of B- to non-B transitions (Sinden, 1994; Majumdar and Patel, 2002; Kouzine *et al.*, 2008; Brooks and Hurley, 2009). Despite this empirical support, the association between supercoiling, triplex formation, and replication arrest remains counterintuitive. This is because progression of a replication complex leads to positive supercoiling ahead of the replication complex (and negative supercoiling behind it), which is expected to disfavor triplex formation. Thus, it is possible that DNA polymerase arrest might be due to either a pre-existing strand break within or near a triplex structure (Aguilera and García-Muse, 2013; León-Ortiz *et al.*, 2014), or to more complex interactions between the parental and daughter strands that may favor inter-strand (rather than intra-strand) triplexes at replication intermediates, such as those noted at precatenanes (Cebrián *et al.*, 2015).

We further speculate that resolution of triplex DNA structures at replication forks might include incision by components of the nucleotide excision repair pathway (Zhao *et al.*, 2009; Kaushik Tiwari and Rogers, 2013), or other pathways, either at one or both ends of the mirror repeats, which represent triplex-to-duplex junctions susceptible to nuclease cleavage (Bacolla and Wu, 1991), or at the single-stranded nucleotides. Incision may be followed either by excision of the nicked strand and resolution of the structure, resulting in a micro-deletion, or by DNA synthesis templated by the mirror repeat fragment, resulting in a micro-insertion (Figure 2). Breakage and excision involving the entire mirror repeat fragment plus the intervening sequence might account for those cases (79 micro-deletions and 8 micro-insertions

in our study) in which the micro-deletion or micro-insertion included longer fragments than the mirror repeats themselves (not shown).

In vitro studies have demonstrated that different DNA helicases, such as RecQ (Bacolla *et al.*, 2011), DHX9 (Jain *et al.*, 2013), and ChlR1 (Guo *et al.*, 2015), are able to resolve triplex DNA structures (León-Ortiz *et al.*, 2014), and that the lack of these activities is generally associated with genomic instability (Bacolla *et al.*, 2011; Jain *et al.*, 2013; Guo *et al.*, 2015). Thus, it is possible that DNA helicase activity might also contribute to the processing of R•Y-rich mirror repeats following the initial single strand break.

Missense and Nonsense Mutations at Direct and Inverted Repeats

Approximately 15% of the 47,119 missense mutations and ~12% of the 12,362 nonsense mutations resulted either in the formation of perfect direct repeats from interrupted repeats, or in the extension of the pre-existing direct repeats (Table 2). These proportions are significantly higher (Fisher's Exact test, $p < 10^{-7}$) than in the corresponding *in silico* generated control datasets, and are potentially explicable in terms of slipped strand mispairing, or non-B DNA slipped structure formation (Figures 3-4) (Iyer *et al.*, 2015). DNA polymerase slippage on either strand of direct repeat sequences may generate mismatches due to misinsertion, particularly when nucleotide addition occurs at the end of the repeat tracts (Mukherjee *et al.*, 2013, 2014; Bacolla *et al.*, 2015).

The fidelity of DNA polymerases is also dependent upon the local DNA sequence architecture and the identity of the DNA polymerase involved (Kunkel and Bebenek, 2000; Kunkel, 2004). Thus, a single base-pair substitution may be established upon misincorporation, either at the subsequent round of replication or by the erroneous removal of the correct base by DNA repair, e.g. mismatch repair (Schofield and Hsieh, 2003; Kunkel and Erie, 2005). If these activities occur at a preexisting interruption along a direct repeat, a longer (uninterrupted) direct

repeat may be generated. By contrast, if these activities take place along an interrupted direct repeat, a shorter (interrupted) direct repeat will be created.

The proportions of missense (~14%) and nonsense (~12%) mutations that resulted in the creation of perfect inverted repeats or to the extension of existing inverted repeats were significantly higher ($p = 2.94 \times 10^{-13}$ and 6.42×10^{-16} , respectively) than expected. This may be explicable in terms of the proposed mechanism of palindrome correction, hairpin loop formation and repair (Figures 5-6).

Missense and Nonsense Mutations at G-Quartets

The number of missense mutations occurring within G-quartets (2,903/47,119) was significantly higher (Table 1; $p=9.22 \times 10^{-156}$) than expected. The vast majority of these mutations (98%; 2,855/2,903) were observed within G-quartets formed by G-runs of length 2. Despite strong evidence for intramolecular G-quadruplex formation in cells (Lam *et al.*, 2013; Murat and Balasubramanian, 2014), the cascade of events leading to the generation of germline single base-pair substitutions at G-quartets remains to be fully elucidated. Mutations occurring within G-quartets may be generated during DNA replication (Lopes *et al.*, 2011; Boyer *et al.*, 2013; van Kregten and Tijsterman, 2014; Wickramasinghe *et al.*, 2015), a time when the number of such structures increases during the S-phase of the cell cycle (Biffi *et al.*, 2013), particularly on the lagging strand template (Bochman *et al.*, 2012). It is conceivable that single base-pair substitutions within G-quadruplexes could destabilize the structures by decreasing stacking, the degree of destabilization depending upon the position of a mutation within G-quartets (Lee and Kim, 2009). Alternatively, or in addition, base substitution at guanines within G-quartets may involve preferential oxidation during transcription, as a result of increased exposure to cellular oxidants while in their non-canonical duplex configuration (Clark *et al.*, 2012; Zhou *et al.*, 2015). This latter model appears to be supported by the observation that, in

mitochondrial DNA which is likely to come into contact with mitochondrial-generated oxidants, deletion breakpoints are observed at high frequencies near G-quartets (Bharti *et al.*, 2014; Dong *et al.*, 2014). The occurrence of a mutation may either follow or precede unwinding of these G4 structures by DNA helicases, such as FANCI (Wu *et al.*, 2008), CHL1 (Wu *et al.*, 2012), PIF1 (Sanders, 2010) or the recently studied ATP-dependent DEAH-box helicase DHX36 (Chen *et al.*, 2015). Thus, complete resolution of the structure, followed by the continuation and completion of DNA replication, could form part of the mutational mechanism.

In summary, our results indicate that direct repeats, inverted repeats and G-quadruplex-forming sequences play a prominent role in mediating missense mutations, whereas only direct repeats and inverted repeats appear to mediate nonsense mutations. This bias reflects the lower probability (<0.21) of finding codons at G-quadruplex-forming motifs that are capable of being converted to a stop codon (TAA, TAG and TGA) by a single nucleotide substitution. The probability was calculated for all such codons (GGA, TGG, gGAG and gGAA) using data reported by Mort *et al.* (2008) and assuming that for codons GAG and GAA at least one nucleotide G is adjacent to these codons (shown as lower case letter g).

Other Factors Influencing Mutability at Non-B DNA-Forming Sequences

The possibility of generating a mutation by one of the mechanisms proposed above is also likely to be influenced by other factors, such as pH, local cation concentration and leading versus lagging replicating DNA strands (Sinden, 1994; Raghavan and Lieber, 2007; Bacolla *et al.*, 2010; Sharma, 2011), all of which are known to play important roles in the kinetics of B- to non-B DNA transitions. These additional factors make the quantitative assessment of the relative role of each non-B DNA structure in mediating micro-lesions uncertain. This notwithstanding, the significantly greater association of direct repeats with micro-lesions, relative to the other types of repeat, may reflect their independence of pH and other parameters.

Indeed, whereas a hairpin structure would only require Na⁺ ions for stabilization, triplex structures may be further stabilized by other factors, such as acidic pH, Mg²⁺ and spermine and spermidine ions (Raghavan and Lieber, 2007; Bacolla *et al.*, 2010; Sharma, 2011) whereas G-quadruplexes are best stabilized by K⁺ ions (Chen and Yang, 2012).

Conclusions

This is the first study to assess the extent to which four different types of non-B DNA-forming sequence – direct, inverted, R•Y-rich mirror repeats and G-quartets – capable of slipped, hairpin/cruciform, triplex and G-quadruplex structure formation, respectively, may be involved in mediating different types of micro-lesion causing (or associated with) human inherited disease. Three novel mechanisms of mutagenesis, based on either the formation or resolution of non-B DNA structures, have been proposed, which together provide new insights into the mutagenesis mediated by non-B DNA-forming sequences.

References

- Aguilera A, García-Muse T. 2013. Causes of genome instability. *Annu Rev Genet* 47:1-32.
- Bacolla A, Cooper DN, Vasquez KM. 2010. Non-B DNA structure and mutations causing human genetic disease. In: *Encyclopedia of Life Sciences (ELS)*. Chichester: John Wiley & Sons, Ltd. DOI: 10.1002/9780470015902.a0022657.
- Bacolla A, Wang G, Jain A, Chuzhanova NA, Cer RZ, Collins JR, Cooper DN, Bohr VA, Vasquez KM. 2011. Non-B DNA-forming sequences and WRN deficiency independently increase the frequency of base substitution in human cells. *J Biol Chem* 286:10017-10026.
- Bacolla A, Wu FY. 1991. Mung bean nuclease cleavage pattern at a polypurine.polypyrimidine sequence upstream from the mouse metallothionein-I gene. *Nucleic Acids Res* 19:1639-1647.

Bacolla A, Zhu X, Chen H, Howells K, Cooper DN, Vasquez KM. 2015. Local DNA dynamics shape mutational patterns of mononucleotide repeats in human genomes. *Nucleic Acids Res* 43:5065-5080.

Ball EV, Stenson PD, Abeysinghe SS, Krawczak M, Cooper DN, Chuzhanova NA. 2005. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum Mutat* 26:205-213.

Belotserkovskii BP, De Silva E, Tornaletti S, Wang G, Vasquez KM, Hanawalt PC. 2007. A triplex-forming sequence from the human c-MYC promoter interferes with DNA transcription. *J Biol Chem* 282:32433-32441.

Bharti SK, Sommers JA, Zhou J, Kaplan DL, Spelbrink JN, Mergny JL, Brosh RM Jr. 2014. DNA sequences proximal to human mitochondrial DNA deletion breakpoints prevalent in human disease form G-quadruplexes, a class of DNA structures inefficiently unwound by the mitochondrial replicative Twinkle helicase. *J Biol Chem* 289:29975-29993.

Biffi G, Tannahill D, McCafferty J, Balasubramanian S. 2013. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat Chem* 5:182-186.

Bochman ML, Paeschke K, Zakian VA. 2012. DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* 13:770-780.

Boyer AS, Grgurevic S, Cazaux C, Hoffmann JS. 2013. The human specialized DNA polymerases and non-B DNA: vital relationships to preserve genome integrity. *J Mol Biol* 425:4767-4781.

Brooks TA, Hurley LH. 2009. The role of supercoiling in transcriptional control of MYC and its importance in molecular therapeutics. *Nat Rev Cancer* 9:849-861.

Bzymek M, Saveson CJ, Feschenko VV, Lovett ST. 1999. Slipped misalignment mechanisms of deletion formation: *in vivo* susceptibility to nucleases. *J Bacteriol* 181:477–482.

Cebrián J, Castán A, Martínez V, Kadomatsu-Hermosa MJ, Parra C, Fernández-Nestosa MJ, Schaerer C, Hernández P, Krimer DB, Schwartzman JB. 2015. Direct evidence for the formation of precatenanes during DNA replication. *J Biol Chem* 290:13725-13735.

Chen Y, Yang D. 2012. Sequence, stability, and structure of G-quadruplexes and their interactions with drugs. In: *Current Protocols in Nucleic Acid Chemistry*. John Wiley & Sons, p 1–26.

Chen MC, Murat P, Abecassis K, Ferré-D'Amaré AR, Balasubramanian S. 2015. Insights into the mechanism of a G-quadruplex-unwinding DEAH-box helicase. *Nucleic Acids Res* 43:2223-2231.

Chuzhanova NA, Chen JM, Bacolla A, Patrinos GP, Férec C, Wells RD, Cooper DN. 2009. Gene conversion causing human inherited disease: the evidence for involvement of recombination-associated motifs and non-B DNA-forming sequences in DNA breakage. *Hum Mutat* 30:1189-1198.

Clark DW, Phang T, Edwards MG, Geraci MW, Gillespie MN. 2012. Promoter G-quadruplex sequences are targets for base oxidation and strand cleavage during hypoxia-induced transcription. *Free Radic Biol Med* 53:51-99.

Cooper DN, Bacolla A, Férec C, Vasquez KM, Kehrer-Sawatzki H, Chen JM. 2011. On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. *Hum Mutat* 32:1075-1099.

Cooper DN, Krawczak M. 1993. *Human Gene Mutation*. Oxford: Bios Scientific. 412 pp.

Cooper DN, Krawczak M. 1991. Mechanisms of insertional mutagenesis in human genes causing genetic disease. *Hum Genet* 87:409-415.

De S, Michor F. 2011. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat Struct Mol Biol* 18:950-955.

Dong DW, Pereira F, Barrett SP, Kolesar JE, Cao K, Damas J, Yatsunyk LA, Johnson FB, Kaufman BA. 2014. Association of G-quadruplex forming sequences with human mtDNA deletion breakpoints. *BMC Genomics* 15:677.

Du X, Gertz EM, Wojtowicz D, Zhabinskaya D, Levens D, Benham CJ, Schäffer AA, Przytycka TM. 2014. Potential non-B DNA regions in the human genome are associated with higher rates of nucleotide mutation and expression variation. *Nucleic Acids Res* 42:12367-12379.

Greenblatt MS, Grollman AP, Harris CC. 1996. Deletions and insertions in the p53 tumor suppressor gene in human cancers: confirmation of the DNA polymerase slippage/misalignment model. *Cancer Res* 56:2130–2136.

Guo M, Hundseth K, Ding H, Vidhyasagar V, Inoue A, Nguyen CH, Zain R, Lee JS, Wu YA. 2015. Distinct triplex DNA unwinding activity of ChlR1 helicase. *J Biol Chem* 290:5174-5189.

Gusev VD, Nemytikova LA, Chuzhanova NA. 1999. On the complexity measures of genetic sequences. *Bioinformatics* 15:994-999.

Iyer RR, Pluciennik A, Napierala M, Wells RD. 2015. DNA triplet repeat expansion and mismatch repair. *Annu Rev Biochem* 84:199-226.

Jain A, Bacolla A, Del Mundo IM, Zhao J, Wang G, Vasquez KM. 2013. DHX9 helicase is involved in preventing genomic instability induced by alternatively structured DNA in human cells. *Nucleic Acids Res* 41:10345-10357.

Javadekar SM, Raghavan SC. 2015. Snaps and mends: DNA breaks and chromosomal translocations. *FEBS J*. doi: 10.1111/febs.13311. [Epub ahead of print].

Jeitany M, Pineda JR, Liu Q, Porreca RM, Hoffschir F, Desmaze C, Silvestre DC, Mailliet P, Junier MP, Londoño-Vallejo A, Ségal-Bendirdjian E, Chneiweiss H, Boussin FD. 2015. A preclinical mouse model of glioma with an alternative mechanism of telomere maintenance (ALT). *Int J Cancer* 136:1546-1558.

Kaushik Tiwari M, Rogers FA. 2013. XPD-dependent activation of apoptosis in response to triplex-induced DNA damage. *Nucleic Acids Res* 41:8979-8994.

Kondrashov AS, Rogozin IB. 2004. Context of deletions and insertions in human coding sequences. *Hum Mutat* 23:177–185.

Kouzine F, Sanford S, Elisha-Feil Z, Levens D. 2008. The functional response of upstream DNA to dynamic supercoiling in vivo. *Nat Struct Mol Biol* 15:146–154.

Kunkel TA. 2004. DNA replication fidelity. *J Biol Chem* 279:16895-16898.

Kunkel TA, Bebenek K. 2000. DNA replication fidelity. *Annu Rev Biochem* 69:497–529.

Kunkel TA, Erie DA. 2005. DNA mismatch repair. *Annu Rev Biochem* 74:681–710.

Lam EY, Beraldi D, Tannahill D, Balasubramanian S. 2013. G-quadruplex structures are stable and detectable in human genomic DNA. *Nat Commun* 4:1796.

Lee JY, Kim DS. 2009. Dramatic effect of single-base mutation on the conformational

dynamics of human telomeric G-quadruplex. *Nucleic Acids Res* 37:3625–3634.

León-Ortiz AM, Svendsen J, Boulton SJ. 2014. Metabolism of DNA secondary structures at the eukaryotic replication fork. *DNA Repair (Amst)* 19:152-162.

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 4:203–221.

Lopes J, Piazza A, Bermejo R, Kriegsmann B, Colosio A, Teulade-Fichou MP, Foiani M, Nicolas A. 2011. G-quadruplex-induced instability during leading-strand replication. *EMBO J* 30:4033-4046.

Lovett ST. 2004. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol Microbiol* 52:1243–1253.

Lu S, Wang G, Bacolla A, Zhao J, Spitser S, Vasquez KM. 2015. Short Inverted Repeats Are Hotspots for Genetic Instability: Relevance to Cancer Genomes. *Cell Rep*. doi: 10.1016/j.celrep.2015.02.039. [Epub ahead of print].

Majumdar A, Patel DJ. 2002. Identifying hydrogen bond alignments in multistranded DNA architectures by NMR. *Acc Chem Res* 35:1–11.

Mort M, Ivanov D, Cooper DN, Chuzhanova NA. 2008. A meta-analysis of nonsense mutations causing human genetic disease. *Hum Mutat* 29:1037-1047.

Mukherjee P, Lahiri I, Pata JD. 2013. Human polymerase kappa uses a template-slippage deletion mechanism, but can realign the slipped strands to favour base substitution mutations over deletions. *Nucleic Acids Res* 41:5024–5035.

Mukherjee P, Wilson RC, Lahiri I, Pata JD. 2014. Three residues of the interdomain linker determine the conformation and single-base deletion fidelity of Y-family translesion polymerases. *J Biol Chem* 289:6323-6331.

Murat P, Balasubramanian S. 2013. Existence and consequences of G-quadruplex structures in DNA. *Curr Opin Genet Dev* 25C:22–29.

Nambiar M, Srivastava M, Gopalakrishnan V, Sankaran SK, Raghavan SC. 2013. G-quadruplex structures formed at the HOX11 breakpoint region contribute to its fragility during t(10;14) translocation in T-cell leukemia. *Mol Cell Biol* 33:4266-4281.

Patel HP, Lu L, Blaszak RT, Bissler JJ. 2004. *PKDI* intron 21: triplex DNA formation and effect on replication. *Nucleic Acids Res* 32:1460–1468.

Raghavan SC, Lieber MR. 2007. DNA structure and human diseases. *Front Biosci* 12:4402-4408.

Rosche WA, Trinh TQ, Sinden RR. 1995. Differential DNA secondary structure-mediated deletion mutation in the leading and lagging strands. *J Bacteriol* 177:4385–4391.

Rouleau SG, Beaudoin J, Bisailon M, Perreault J. 2015. Small antisense oligonucleotides against G-quadruplexes: specific mRNA translational switches. *Nucleic Acids Res* 43:595-606.

Sanders CM. 2010. Human Pif1 helicase is a G-quadruplex DNA-binding protein with G-quadruplex DNA-unwinding activity. *Biochem J* 430:119–128.

Sen D, Gilbert W. 1988. Formation of parallel four-stranded complexes by guanine rich motifs in DNA and its implications for meiosis. *Nature* 334:364–366.

Sharma S. 2011. Non-B DNA secondary structures and their resolution by RecQ helicases. *J Nucleic Acids* 2011:724215.

Sinden RR. 1994. *DNA structure and function*. Academic Press. 394pp.

Sinden RR, Wells RD. 1992. DNA structure, mutations, and human genetic disease. *Curr Opin Biotechnol* 3:612–622.

Spurlock G, Bennett E, Chuzhanova N, Thomas N, Jim HP, Side L, Davies S, Haan E, Kerr B, Huson SM, Upadhyaya M. 2009. *SPRED1* mutations (Legius syndrome): another clinically useful genotype for dissecting the neurofibromatosis type 1 phenotype. *J Med Genet* 46:431-437.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The Human Gene Mutation Database: 2008 update. *Genome Med* 1:13.

Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. 2014. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 133:1-9.

Tappino B, Chuzhanova NA, Regis S, Dardis A, Corsolini F, Stroppiano M, Tonoli E, Beccari T, Rosano C, Mucha J, Blanco M, Szlago M, Di Rocco M, Cooper DN, Filocamo M. 2009. Molecular characterization of 22 novel UDP-N-acetylglucosamine-1-phosphate transferase α - and β -subunit (*GNPTAB*) gene mutations causing mucopolipidosis types II α / β and III α / β in 46 patients. *Hum Mutat* 30:E956–E973.

van Kregten M, Tijsterman M. 2014. The repair of G-quadruplex-induced DNA damage. *Exp Cell Res* 329:178-183.

Vasquez KM, Wang G. 2013. The yin and yang of repair mechanisms in DNA structure-

induced genetic instability. *Mutat Res* 743:118–131.

Verdin H, D'haene B, Beysen D, Novikova Y, Menten B, Sante T, Lapunzina P, Nevado J, Carvalho CM, Lupski JR, De Baere E. 2013. Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the *FOXL2* gene or its regulatory domain. *PLoS Genet* 9:e1003358.

Wells RD. 2007. Non-B DNA conformations, mutagenesis and disease. *Trends Biochem Sci* 32:271-278.

Wang G, Vasquez KM. 2014. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair (Amst)* 19:143-151.

Wickramasinghe CM, Arzouk H, Frey A, Maiter A, Sale JE. 2015. Contributions of the specialised DNA polymerases to replication of structured DNA. *DNA Repair (Amst)* 29:83-90.

Williams JD, Fleetwood S, Berroyer A, Kim N, Larson ED. 2015. Sites of instability in the human *TCF3 (E2A)* gene adopt G-quadruplex DNA structures *in vitro*. *Front Genet* 6:177.

Wu Y, Shin-Ya K, Brosh RM. 2008. FANCI helicase defective in Fanconi anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Mol Cell Biol* 28:4116–4128.

Wu Y, Sommers JA, Khan I, de Winter JP, Brosh RM. 2012. Biochemical characterization of Warsaw breakage syndrome helicase. *J Biol Chem* 287:1007–1021.

Wu X, Lu Y, Ding Q, You G, Dai J, Xi X, Wang H, Wang X. 2014. Characterisation of large F9 deletions in seven unrelated patients with severe haemophilia B. *Thromb Haemost* 112:459-465.

You GL, Ding QL, Lu YL, Dai J, Xi XD, Wang XF, Wang HL. 2013. Characterization of large deletions in the *F8* gene using multiple competitive amplification and the genome walking technique. *J Thromb Haemost* 11:1103-1110.

Zhao J, Jain A, Iyer RR, Modrich PL, Vasquez KM. 2009. Mismatch repair and nucleotide excision repair proteins cooperate in the recognition of DNA interstrand crosslinks. *Nucleic Acids Res* 37:4420-4429.

Zhao XN, Usdin K. 2015. The repeat expansion diseases: the dark side of DNA repair. *DNA Repair (Amst)*. doi: 10.1016/j.dnarep.2015.04.019. [Epub ahead of print].

Zhou J, Fleming AM, Averill AM, Burrows CJ, Wallace SS. 2015. The NEIL glycosylases remove oxidized guanine lesions from telomeric and promoter quadruplex DNA structures. *Nucleic Acids Res* 43:4039-4054.

Legends to Figures

Figure 1. Examples of non-B DNA-forming repeats and their corresponding secondary structures. Only one of several possible conformers is shown for clarity.

Figure 2. Schematic representation of a putative mechanism accounting for micro-deletions (steps A – B – C – D - E) and micro-insertions (steps A - B - C' - D' - E') mediated by triplex structure formation at R•Y-rich mirror repeat sequences.

Figure 3. Schematic representation of putative mechanisms leading to single base-pair substitutions mediated by the formation of perfect direct repeats from imperfect repeat sequences.

Figure 4. Schematic representation of putative mechanisms capable of explaining single base-pair substitutions mediated by the extension of perfect direct repeats.

Figure 5. Schematic representation of potential mechanisms accounting for single base-pair substitutions mediated by the formation of perfect inverted repeats from imperfect repeat sequences.

Figure 6. Putative mechanisms for the generation of single base-pair substitutions through the extension of existing perfect inverted repeats.

Table 1. Mutations and breakpoints overrepresented in direct, inverted, R•Y-rich mirror repeats, and in G-quadruplexes.

Mutation/ breakpoint type	Repeat type	Dataset	Number of mutations		% mutations in-repeat	p-value
			in-repeat	not-in-repeat		
Micro-deletions first breakpoints	Direct	HGMD ^a	3612	13298	21.36	1.54×10^{-8} Over ^c
		Expected ^b	3313282	13596718	19.59	
	Inverted	HGMD	2563	14347	15.16	0.032 Over
		Expected	2452172	14457828	14.5	
	Mirror	HGMD	870	16040	5.14	3.56×10^{-41} Over
		Expected	556339	16353661	3.29	
Micro-insertions	Direct	HGMD	1495	5567	21.17	4.8×10^{-5} Over
		Expected	1354492	5707508	19.18	
	Mirror	HGMD	318	6744	4.5	6.04×10^{-11} Over
		Expected	220336	6841664	3.12	
Missense mutations	G- quartet	HGMD	2903	44216	6.16	9.22×10^{-156} Over
		Expected	178945	4532955	3.79	

^aMutations recorded in the HGMD dataset (observed mutations); ^b*in silico* generated random mutations (expected mutations); ^c Over, indicates that the number of observed mutations in the corresponding repeat type is significantly overrepresented after Bonferroni correction for multiple testing.

Table 2. Number of missense and nonsense mutations creating perfect repeats from imperfect repeats.

Mutations	Repeat type	Dataset	Number of perfect repeats		% created	p-value
			created	not created		
Missense	Direct	HGMD ^a	7034	40133	14.91	1.88 × 10 ⁻⁸ Over ^c
		Expected ^b	659195	4052705	13.99	
	Inverted	HGMD	6528	40639	13.84	2.94 × 10 ⁻¹³ Over
		Expected	598412	4113488	12.70	
	Mirror	HGMD	6157	41010	13.05	0.392 NS ^d
		Expected	613019	4098881	13.01	
Nonsense	Direct	HGMD	1538	10824	12.44	2.09 × 10 ⁻⁹ Over
		Expected	132633	1103567	10.72	
	Inverted	HGMD	1474	10888	11.92	6.42 × 10 ⁻¹⁶ Over
		Expected	120283	1115917	9.73	
	Mirror	HGMD	1555	10807	12.57	0.482 NS
		Expected	155713	1080487	12.59	

^aMutations in the HGMD dataset (observed mutations); ^b*in silico* generated random mutations (expected mutations); ^cOver, indicates that the number of observed mutations in the corresponding repeat type is significantly overrepresented after Bonferroni correction for multiple testing; ^dNS, indicates that the result was not significant before correction for multiple testing.

Figure 1

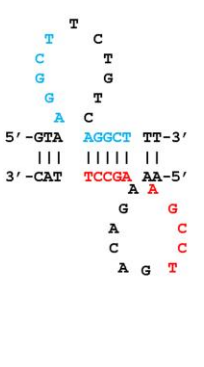
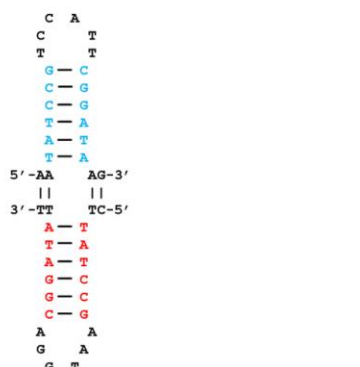
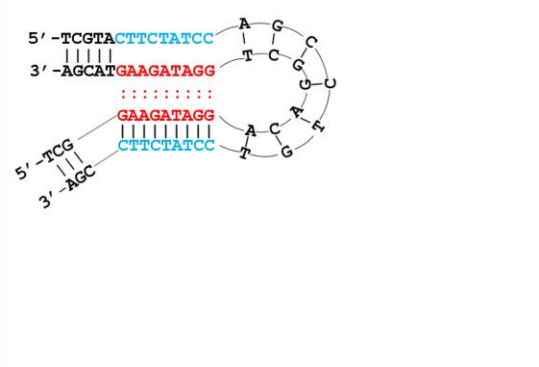
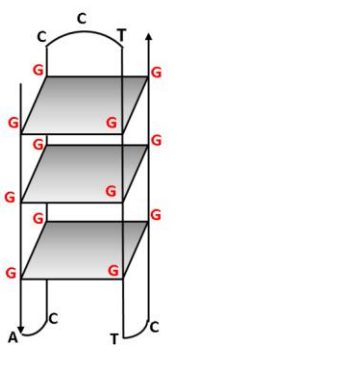
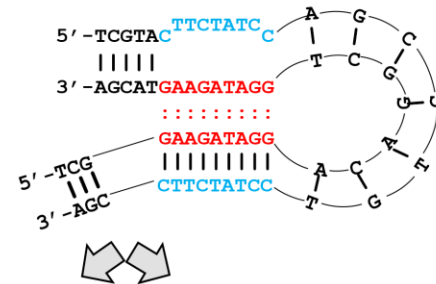
Non-B DNA structures	Slipped structure	Cruciform	Triplex	G-quadruplex
<p>Example of repeat</p>	<p>Direct repeat</p> <p>Example:</p> <p>5' -GTAAGGCTTCTGTCAGGCTTT-3'</p> <p> </p> <p>3' -CATTCCGAAAGACAGTCCGAAA-5'</p>	<p>Inverted repeat</p> <p>Example:</p> <p>5' -AATATCCGTCCATTCGGATAAAG-3'</p> <p> </p> <p>3' -TTATAGGCAGGTAAGCCTATTC-5'</p>	<p>Symmetric (mirror) repeat</p> <p>Example:</p> <p>5' -TCGTACTTCTATCCAGCCTGTCCTATCTTCCGA-3'</p> <p> </p> <p>3' -AGCATGAAGATAGGTCGGACAGGATAGAAGGCT-5'</p>	<p>G-quartet</p> <p>Example:</p> <p>5' -GGGACGGGCCTGGGTCGGG-3'</p>
<p>Secondary structure</p>				

Figure 2

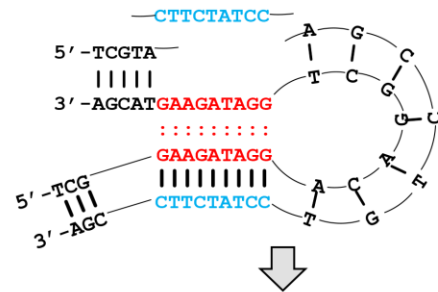
Step A. Fragment of DNA strands containing mirror repeats



Step B. Formation of non-B triplex structure at the mirror repeat symmetry



Step C. Excision of the single-stranded mirror repeat (cyan)



Step D. Secondary structure resolution

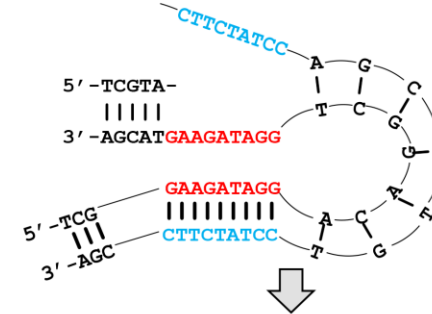


Step E. Cleavage and repair of the unpaired mirror repeat

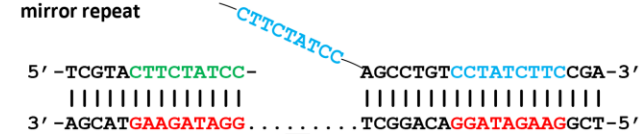


Micro-deletion

Step C'. Incision at the upstream duplex/single-stranded junction



Step D'. Structure resolution and first DNA synthesis across the unpaired mirror repeat



Step E'. Incision opposite the flap, second DNA synthesis across the unpaired "flap" and "repair"



Micro-insertion

Figure 3

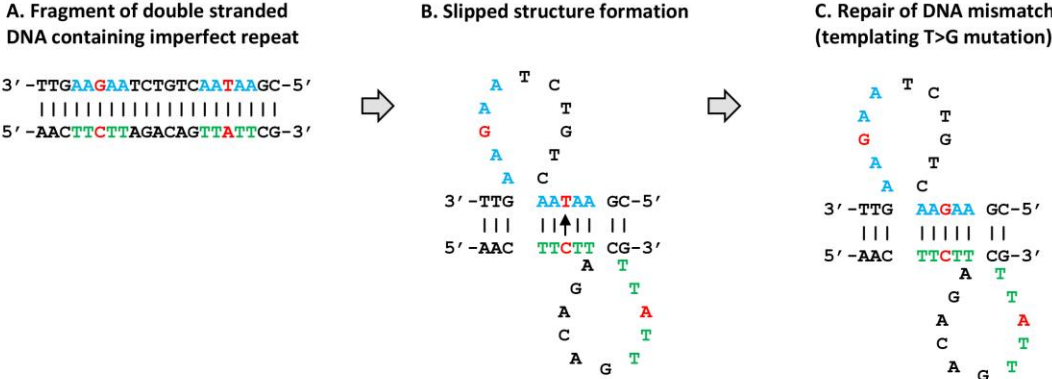


Figure 4

A. Fragment of double stranded DNA containing perfect repeat

3' -ACGACCAATGTCCGACCA-5'
|||||
5' -TGCTGGTTACAGGCTGGGT-3'



B. Slipped structure formation

 C A A
 C T
 A G
 G T
 C C
3' -A CGACCC A-5'
 | | | | |
5' -T GCTGGT T-3'
 T G
 A G
 C G
 A T
 G G C



C. C>A mutation and extension of the existing repeat

 C A A
 C T
 A G
 G T
 C C
3' -A CGACCA A-5'
 | | | | |
5' -T GCTGGT T-3'
 T G
 A G
 C G
 A T
 G G C

Figure 5

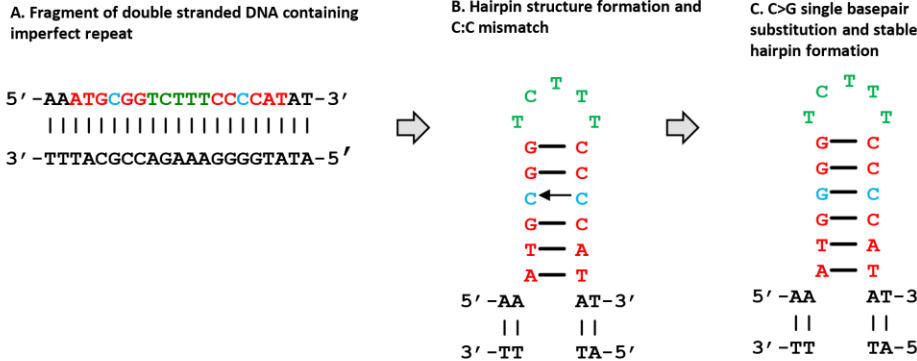


Figure 6

