Cell. Mol. Life Sci. (2010) 67:43–62
DOI 10.1007/s00018-009-0131-2

Cellular and Molecular Life Sciences

REVIEW

Non-B DNA structure-induced genetic instability and evolution

Junhua Zhao · Albino Bacolla · Guliang Wang · Karen M. Vasquez

Received: 18 June 2009 / Revised: 22 July 2009 / Accepted: 11 August 2009 / Published online: 1 September 2009

Abstract Repetitive DNA motifs are abundant in the genomes of various species and have the capacity to adopt non-canonical (i.e., non-B) DNA structures. Several non-B DNA structures, including cruciforms, slipped structures, triplexes, G-quadruplexes, and Z-DNA, have been shown to cause mutations, such as deletions, expansions, and translocations in both prokaryotes and eukaryotes. Their distributions in genomes are not random and often co-localize with sites of chromosomal breakage associated with genetic diseases. Current genome-wide sequence analyses suggest that the genomic instabilities induced by non-B DNA structure-forming sequences not only result in predisposition to disease, but also contribute to rapid evolutionary changes, particularly in genes associated with development and regulatory functions. In this review, we describe the occurrence of non-B DNA-forming sequences in various species, the classes of genes enriched in non-B DNA-forming sequences, and recent mechanistic studies on DNA structure-induced genomic instability to highlight their importance in genomes.

Keywords Non-B DNA structures · Repetitive sequences · Genomic instability · Evolutionary change · Gene regulation

Electronic supplementary material The online version of this article (doi:10.1007/s00018-009-0131-2) contains supplementary material, which is available to authorized users.

J. Zhao · A. Bacolla · G. Wang · K. M. Vasquez (✉)
Department of Carcinogenesis, Science Park-Research Division,
The University of Texas M.D. Anderson Cancer Center,
1808 Park Road 1-C, P.O. Box 389, Smithville, TX 78957, USA
e-mail: kvasquez@mdanderson.org

Introduction

The canonical right-handed double helical structure of B-form DNA [1] has had a profound influence over studies designed to determine the function of DNA. However, many alternative DNA structures [2] have been known to exist since the late 1950s and their roles in biological functions have begun to be elucidated, with substantial progress over the past decade. In 1957, sedimentation coefficient and optical absorption measurements revealed the association of ribonucleic poly-A and poly-U polymers into three-stranded complexes [3]. The DNA of a d(CpGpCpGpCpG) fragment was crystallized in 1979, which revealed a left-handed conformation (Z-DNA) with altered helical parameters relative to the right-handed B-form [4]. Soon after, cruciform structures formed by inverted repeats were identified by S1 nuclease probing [5, 6] and by two-dimensional gel electrophoresis [7]. During this same period, parallel four-stranded complexes (tetraplex or G-quadruplex DNA) were discovered to form by guanine-rich DNA sequences [8]. To date, more than ten different DNA conformations are known to exist from biophysical and biochemical studies, and more are likely to be identified.

Non-B DNA-forming sequences in genomes affect DNA replication and transcription, and contribute to genome instability [9–12]. In 1984, Glickman and Ripley [13] reported the induction of deletions in the *lacI* gene of *Escherichia coli* by putative cruciform structures. More recently, studies in model systems (bacteria, yeast, and mammalian cell culture) on trinucleotide repeat sequences, whose expansions in disease-related genes are involved in approximately 30 human hereditary neurological disorders [14], support the mutagenic role of non-B DNA structures [15–17]. Similarly, DNA sequences, capable of forming non-canonical structures from the human *c-MYC* and *BCL-2* loci that co-localize with translocation breakpoints, undergo frequent double-strand breaks (DSBs) in mammalian cells [12, 18–22]. In support of these results, the same non-B DNA structure-forming sequence from the human *c-MYC* gene also stimulates genomic instability on chromosomes in transgenic mice [23]. Large (A + T)-rich inverted repeats on chromosome 22q11 and other chromosomes, such as 11q23 and 17q11, were found to cause recurrent translocations both in sperm cells in the general population [24] and in cell culture [25], providing evidence that cruciform structures may cause genomic rearrangements [26–28]. Thus, alternative DNA conformations are believed to contribute to mutations and to the dysregulation of cancer-related genes in translocation-related malignant diseases such as myeloma, leukemia, and lymphoma [11, 29].

Recent advances in the field of genomics have revealed the widespread occurrence of non-B DNA-forming motifs in various genomes, their selective enrichment within specific classes of genes and/or chromosomes, and the asymmetric frequency distributions within transcriptional units. These data are paradoxical given the mutagenic role of repeating sequences and their involvement in human disease. Herein, we describe the structural features and biological functions (and potential mechanisms involved) of the most well-characterized DNA structures, i.e., Z-DNA, cruciforms, triplexes or H-DNA, G-quadruplexes, and looped-out slipped structures. We also provide evidence for novel roles of non-B DNA structure-forming sequences as co-regulators of transcriptional activity and as genomic elements through which positive selective pressures have acted during evolutionary time so as to shape and preserve specific genomic functions.

Non-B DNA structures

The distribution of nucleotides in genomes is not random. Many DNA sequence patterns exist throughout genomes from bacteria to human, such as direct repeats of homo-, di-, or tri-nucleotides, inverted repeats, mirror repeats, etc. Unlike the majority of DNA sequences which form the canonical right-handed B-form [1], repeated sequences have the capacity to also adopt alternative conformations (i.e., non-B DNA structures). To date, nearly a dozen types of non-B DNA structures have been described, including hairpins/cruciforms, Z-DNA, triplexes (H-DNA), tetraplexes, slipped DNA, and sticky DNA.

Hairpins/cruciforms

Hairpin/cruciform structures can form at inverted repeats [30]. One side of an inverted repeat, equidistant to the symmetric center, is complementary to the sequence on the other side, e.g., 5′-GACTGC….GCAGTC-3′ (Fig. 1a). The two inverted repeats base-pair with one another and form an intrastrand hairpin stem, leaving the sequence at the symmetric center looped out as a single strand. The cruciform structure consists of two hairpin-loop arms and a 4-way junction, which is structurally similar to a Holliday junction recombination intermediate [31]. Formation of hairpin/cruciform structures from double-stranded DNA requires energy that may come from negative supercoiling [32, 33]. Under these conditions, two inverted repeats as short as 7 bp are sufficient for the formation of hairpin structures [34].

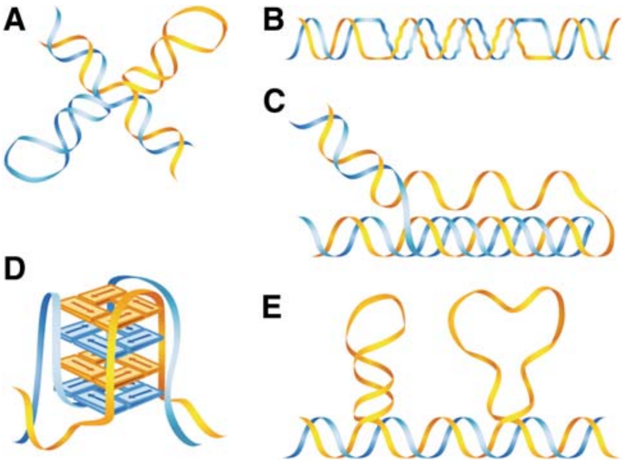

Fig. 1 Non-B DNA structures. **a** Cruciform DNA, **b** Z-DNA, **c** H-DNA (triplex DNA), **d** G-quadruplex (tetraplex) DNA, and **e** slipped DNA

Z-DNA

Sequences with alternating pyrimidines and purines, such as $(CG:CG)_n$ and $(CA:TG)_n$, may wind the double helix into a left-handed zigzag pattern (Z-DNA), as depicted in Fig. 1b. Whereas CG:CG repeats are most likely to form the Z-DNA structure, GT:AC repeats are more abundant in the human genome [35]. Compared to the right-handed B-form, the left-handed Z-DNA contains inverted purines in the *syn*-conformation while pyrimidines remain in the *anti*-conformation with the sugar-pucker altered from the C2- to the C3-endo position so as to maintain the Watson–Crick base-pairing [36]. These alterations cause a change in the sugar–phosphate backbone that changes the organization of the double helix. Therefore, unlike B-form DNA, which possesses one major groove and one minor groove, Z-DNA has only one deep and narrow groove with 12 bp per helical turn [4, 37, 38]. The crystal structure of a B- to Z-DNA junction was solved in 2005 and revealed an extruded base pair on each side of the DNA duplex, which is susceptible to DNA modification [39].

Triplex DNA (H-DNA)

Intramolecular triplex DNA structures can form at homopurine:homopyrimidine sequences with mirror symmetry, where a single-stranded region can bind in the major groove of the underlying DNA duplex to form a three-stranded helix [40–42] (Fig. 1c). Triplex DNA can be classified according to the orientation and composition of the third strand, which can form either Hoogsteen or reverse-Hoogsteen hydrogen bonds with the purine-rich strand of the duplex DNA. Hence, the third strand can be either pyrimidine-rich and parallel to the complementary strand (Y*R:Y), or purine-rich and anti-parallel to the complementary strand (R*R:Y). Whereas (R*R:Y) triplexes form under conditions of physiological pH, triplex structures of the (Y*R:Y) composition form most readily under conditions of acidic pH. At physiological pH, triplex structures may be stabilized by negative supercoiling, modification with phosphorothioate groups, or polyvalent cations such as spermine and spermidine [42].

Tetraplex DNA (G-quadruplex DNA)

This four-stranded structure consists of a square co-planar array of four guanines formed by a stretch of guanine-rich DNA [43] (Fig. 1d). Each guanine acts as a donor and acceptor of Hoogsteen hydrogen bonds in a cyclic arrangement involving N-1, N-2, O-6, and N-7. In vitro, these structures are stabilized by K^+ or Na^+ ions at physiological pH and temperature. Quadruplex structures may be formed by one, two, or four interacting strands and exist in a variety of conformations depending on the polarity of the strands (parallel or anti-parallel), glycosidic torsion angles, groove size, base sequence of the connecting loops, and the participation of cations.

Slipped strand DNA

When direct repeats are base-paired with the complementary strand in a misaligned fashion, a slipped structure forms, particularly following unwinding, yielding hairpins or looped-out bases [44] (Fig. 1e). When direct repeats involve several units, like the triplet repeat sequences (CGG, CTG, and CAG), the looped-out bases may form duplexes stabilized by interstrand stacking interactions [45].

Non-B DNA conformations in vivo

The formation of DNA secondary structures in vitro has been demonstrated by several methods, including polyacrylamide gel electrophoresis, nuclease cleavage, chemical probing, circular dichroism, NMR, ultraviolet absorption, electron microscopy, atomic force microscopy, and crystallography [46].

In vivo, non-B DNA conformations are believed to form, at least transiently, during DNA metabolic processes such as replication, transcription, repair, or recombination [9, 11]. The expansion of slipped DNA-forming trinucleotide repeats observed in neurological diseases [14] correlates with the stability of secondary structures in vitro. For example, interruptions in the trinucleotide repeats of the *SCA1* (CAG:CTG) and *FRAXA* (CCG:CGG) genes exert a protective role against instability [47–49]. These interruptions reduce the propensity of DNA secondary structure formation in vitro, and the correlation between rates of expansion in individuals and slipped DNA formation has been taken to support a role for slipped DNA in genetic instability. Nevertheless, the transient existence of these, and other non-B DNA structures, has made their detection difficult in genomic DNA [50], particularly in cases such as simple repeats, in which multiple conformations are possible, depending upon the environmental conditions [51].

To date, fluorescence immunostaining by antibodies against specific DNA structures rather than the sequences per se is considered the most direct method for detecting non-B DNA structures in vivo. Rabbit antibodies specific for the Z-DNA structure formed by brominated poly [d(GC)]:poly[d(GC)] were generated in 1981 [52], and used to bind the interband regions of *Drosophila* polytene chromosomes [53] and to detect Z-DNA formed by GT repeats in negatively supercoiled plasmids in vitro [54]. Currently, antibodies against Z-DNA are commercially available (Abcam, GeneTex, etc.). One caveat of this methodology is that the estimation of non-B DNA structure formation in vivo may not reflect the physiological equilibrium conditions, since binding of Z-DNA antibodies may shift the B- to Z-DNA equilibrium towards the Z conformation [55].

Several mouse monoclonal antibodies were developed to detect triplex DNA in chromosomes [56, 57], and were demonstrated to bind triplex DNA specifically [58]. H-DNA structures in human interphase nuclei were also detected by fluorescently labeled single-stranded DNA oligonucleotides (complementary to the single-stranded region of the H-DNA structure) in vivo [59]. A quadruplex monoclonal antibody was first developed in 1998 [60, 61] in mice against the quadruplexes formed by synthetic d($CGCG_4$ GCG) and the telomere-derived d(TG_4) and d(T_2G_4)$_4$ sequences in vitro. Later, a high affinity ($K_d \sim 4$ nM) antibody against tetraplex DNA structures was developed and used in in vivo studies of telomeric tetraplex structures in the macronucleus of *Stylonychia lemnae* [62]. Finally, a monoclonal antibody was developed to recognize cruciform and T-shaped DNA structures [63].

Additional evidence for the existence of non-B DNA structures in vivo has been generated using methods such as chemical probing and DNA cross-linking of genomic DNA sequences [64, 65]. However, most of these methods require DNA extraction (before and/or after treatment) for analyses, as it has proven difficult to directly detect these structures in living cells. In addition to technical challenges associated with the detection of non-B DNA, these structures are certainly transient in nature in cells, making their detection even more challenging.

Genome-wide analyses and evolutionary relationships

Abundance and distribution of non-B DNA-forming sequences

Since the abundance and distribution of non-B DNA-forming sequences may provide insights into their functions in DNA metabolism, analyses were carried out to compare the abundance of these structures in the genomes of various organisms. Overall, non-B DNA-forming sequences are more abundant in eukaryotic genomes than in prokaryotes [66].

Hairpin/cruciform (Inverted repeats)

Analysis of human sequences containing 157 genes for a total of 1 Mb of genomic sequence (including exons, introns, and 5′- or 3′-UTRs) revealed many dA:dT sequences, which may form cruciforms [51]. In this sample set, the overall dA:dT abundance was ~49.7%, and the cruciform-forming sequences [≥8-bp (A + T)-rich inverted repeats] in the human genome was ~1/41,700 bp. Additional analyses of genomic sequences in *E. coli* and yeast revealed that the cruciform-forming sequences were more abundant in yeast (1/19,700 bp) and human than in *E. coli* [51]. The distributions of hairpin/cruciform structure-forming sequences often overlap with chromosomal regions prone to gross rearrangements both in somatic and in germ cells [67–69].

Z-DNA

Although the human genome is less (G + C)-rich than prokaryotic genomes, Z-DNA-forming sequences are in fact very abundant. The GT:AC repeats are estimated to account for more than 0.25% of the entire human genome [35]. A computer-based thermodynamic search strategy (Z-Hunt-II) used by the Ho group to analyze the complete human genome showed that Z-DNA-forming sequences occur approximately once every 3,000 bp [70]. Furthermore, Z-DNA-forming regions were found to be distinctly located near the 5′-ends of genes in the genome, and the proximity between these regions and the transcription start sites (TSS) became more pronounced during the divergence from prokaryotes to eukaryotes [70]. Therefore, the location bias of these GT:AC repeats is supportive of Z-DNA formation and stabilization by the transient surges in negative supercoiling associated with transcription. As early as 1983, Nordheim and Rich [71] suggested that three 8-bp Z-DNA-forming sequences in the simian virus 40 enhancer region may function in transcriptional activation. Studies in yeast showed that Z-DNA structures can be induced or stabilized by Z-DNA-binding proteins and function in gene regulation and chromatin-remodeling [72, 73]. The occurrence of Z-DNA-forming sequences at chromosomal breakpoints in human tumors suggests that Z-DNA plays a role in causing genomic instability, perhaps by inducing double-strand breaks and large deletions [18].

H-DNA-forming sequences (R:Y tracts with mirror symmetry)

H-DNA-forming sequences occur at higher levels than expected in mammalian genomes. Using the same 1-Mb sequence sample set from the human genome as in the study of hairpin/cruciform structure-forming sequences, Schroth and Ho [51] found that the occurrence of H-DNA sequences [≥10-bp 100% homopurine:homopyrimidines but <80% (A + T)-rich] in the human genome was ~1/49,400 bp. The distribution of long (≥100 bp) homopurine:homopyrimide sequences in human genes was confined to introns of genes coding for products localized to the cell membrane, phosphorylation, signal transduction, and development and morphogenesis [74]. H-DNA structure-forming sequences are also found flanking proto-oncogenes, such as *c-MYC*, and may cause genomic instability, such as deletions and other rearrangements [12, 23].

Tetraplex (G-quadruplex DNA)

Two independent genome-wide surveys for potential intramolecular G-quadruplex-forming sequences identified ~37,000 sites in the human genome, approximately 1 tetraplex every 10 kb [75, 76], with ~60% of them located outside coding regions [75]. Tetraplex-forming guanine-rich sequences are found in immunoglobulin switch regions [8], telomeric DNA [77, 78], poly (dG) runs [79], and promoter regions [80]. An analysis of promoter regions of 19,268 validated human genes in ENSEMBL (NCBI 34) showed that ~42.7% of human gene promoters contain at least one quadruplex-forming sequence [80]. Du et al. [81, 82] analyzed 13,276 human reference sequence (RefSeq) genes and 2,892 chicken RefSeq genes for potential

G-quadruplex-forming sequences and identified one or more G4 DNA motifs in >60% of the genes studied. The distribution of the more stable form of G-tetraplex, which contains single-nucleotide loops, is more abundant near transcription start sites, suggesting that this stable secondary structure may have been under positive selection to influence the transcription of particular groups of genes [80]. In addition, a high proportion of genes also contain G4 motifs in 3′-UTRs, implying a role in facilitating transcriptional termination, perhaps by weakening the association of an RNA polymerase complex with template DNA [83]. Therefore, the distribution of G-rich sequences in genomes supports their involvement in the regulation of transcription, in addition to other roles, such as homologous recombination [8, 84] and telomere maintenance [78].

Slipped DNA (S-DNA)

Repetitive DNA sequences account for nearly 30% of the human genome, and are interspersed throughout chromosomes [85, 86]. These repeats are referred to as microsatellites (1–7 nt, [48]) or minisatellites (10–100 nt, [87]). Various human diseases have been demonstrated to be associated with either expansion or contraction of microsatellites and minisatellites [48, 87]. Although microsatellites are abundant in the human genome, their representation varies greatly depending on sequence composition. For example, whereas >16,000 tracts comprised of A or T mononucleotide runs were present in the hg16 assembly at length ≥30 nt, only 7 analogous tracts of Gs and Cs were found [88]. Closer examination of the physical properties of tri- and tetra-nucleotide repeats revealed an inverse relationship between their number in vertebrate genomes and the propensity to fold into the hairpin or quadruplex structures [89]. These data suggest that sequences with the propensity to form stable secondary structures have not been maintained as efficiently as their less stable counterparts during evolutionary time. Nevertheless, a comparison of the distribution of these tri- and tetra-nucleotide sequences in protein coding versus noncoding regions revealed that the number of certain "strong secondary structure-forming" sequences, such as AGC, CCG, CCCG, AGCG, CCGG, and ACCG was higher than expected in coding regions [89], supporting the idea that selective pressures acted so as to preserve the amino acid coding ability of these inherently unstable sequences.

It is important to point out that not all the repeated sequences analyzed to date have the same capacity to form non-B DNA structures. The search criteria used in different reports were set to answer different questions. For example, the Ho group alerted that, although (G + C)-rich sequences are abundant in *E. coli*, not all of them meet the requirement for forming stable secondary structures. Rather, these (G + C)-rich repeats in bacteria are mostly recognized as transcription termination sequences when transcribed into RNA [70]. Also, the most abundant tetraplex-forming G-rich sequences in the human genome analyzed by Huppert and Balasubramanian [76] are located on the coding strand and therefore may fold into alternative structures in the RNA transcripts rather than in genomic DNA. Therefore, all repeat-based analyses should be interpreted with the realization that some of these 'unusual' sequences may not form 'unusual' DNA structures.

Gene categories

The completion of the Human Genome Project (HGP) [35, 90] has made it possible to address the question of the distribution of non-B DNA-forming sequences in relation to transcribed DNA. More than 99% of euchromatic DNA, which contains genes and putative genes, is currently assembled. The remaining 0.5–1% of gapped DNA (~24 Mb) mostly contains segmental duplications, i.e., nearly identical sequences present at different chromosomal locations [91], for which clones are available to enable covering. Hence, the data summarized below is expected to capture most of the global genomic organization of genes in relation to non-B DNA-forming sequences. One notable exception is represented by the 18S- and 28S-ribosomal RNA gene arrays in acrocentric chromosomes, which, like centromeric, pericentromeric, and subtelomeric heterochromatin, were not targeted for sequencing. Indeed, few clones are available for such recalcitrant regions. Heterochromatin, which amounts to ~5–7% (~200 Mb) [91], is almost entirely populated by tandem repeats and shows limited transcriptional activity.

The first genome-wide search for inverted repeats (IRs) in the human genome revealed the prevalence of large IRs (96 with arm size ≥8 kb and ≥95% sequence identity) on the X (~25%) and Y (~15%) chromosomes [69]. Of the 49 IRs whose arms shared >97–99% sequence identity, 11 from chromosome X, 6 from chromosome Y, and 1 from chromosome 11 contained genes/gene clusters predominantly expressed in the testis (Table 1). Indeed, all annotated genes present on the IRs from chromosome Y display testis-restricted expression and have a function in sperm production and maturation [92].

A subsequent search for the distribution of long, i.e., ≥100 and ≥250 nt, R:Y tracts within human genes indicated the presence of such sequences in the introns of 1,951 and 228, respectively, non-redundant transcriptional units [74]. Strong enrichment (P values as low as 10^{-15}) was observed for sequences in genes encoding proteins with ion channel activity, cell adhesion, and cell–cell communication functions, particularly in subcellular structures, such as

Table 1 Gene categories and DNA repeats

Gene/gene families in the largest (>100 kb) inverted repeats (IR)

Chromosome	IR arm size (kb)	Gene/gene class	Tissues with predominant expression
Y palindrome P1	1,450.0	DAZ	Testes
Y palindrome P5	495.5	CDY	Testes
Y palindrome P3	283.0	PRY	Testes
Y palindrome P4	190.2	HSFY	Testes
Xp11.22	142.2	GAGE-D2,3	Testes
Xq22.1	140.6	NXF2	Testes
Y palindrome P2	122.0	DAZ	Testes
Xq13.1	119.3	DMRTC1	Testes, kidney, pancreas
11q14.3	103.9	RNF18	Testes, kidney, spleen

Purine:pyrimidine tracts in introns of genes

Gene category/function	*P* value	
	≥250 nt (228 genes)	≥100 nt (1,951 genes)
Ion channel activity	1.95E-05	5.92E-09
Protein binding	3.14E-03	6.25E-15
Glutamate receptor activity	6.11E-04	1.92E-07
Cell adhesion	1.11E-04	3.36E-12
Cell communication	2.19E-04	5.24E-15
Transmission of nerve impulse	1.83E-04	5.24E-08
Synapse	2.18E-02	7.69E-05
Alternative splicing	ND	2E-82
Chromosomal translocations	ND	1E-07

Tetranucleotide repeats (TR) in introns of genes

Gene category/function/attribute	*P* value	
Localization to the membrane	1E-07–5E-30	(Range for 10 gene groups containing: groups 1–8, 8–15 TR units; group 9, 16 and 17 TR units; group 10, ≥18 TR units; 190–1,423 genes/group)
Ion channel	5E-02–1E-13	
Cell adhesion	8E-04–2E-37	
DNA alternative splicing	1E-64	≥8 TR units (4,182 genes)
Chromosomal translocations	2E-07	≥8 TR units (4,182 genes)

Micro/minisatellites (2–11 nt repeats) in cDNAs

Gene category/function	*P* value (coding plus non-coding exons) (2,626 genes)
Transcription regulator activity	2.0E-40
Regulation of cellular processes	2.3E-38
Protein binding	2.0E-33
Sequence-specific DNA binding	3.8E-23
Nuclear localization	9.3E-22
RNA pol. II transcription factor activity	1.2E-16
Axon guidance	2.3E-05
MAPK signaling pathway	2.1E-04
WNT signaling pathway	2.4E-04

G-quadruplex in both 5′- and 3′-UTR

Gene category/function	*P* value	
	5′-UTR	3′-UTR
Guanyl-nt exchange factor activity	7.9E-13	6.3E-12
Rho guanyl-nt exchange factor activity	7.9E-10	1.6E-09
Regulation of Rho signal transduction	2.0E-10	1.6E-09
Transcription factor activity	6.3E-05	3.2E-10
Sequence-specific DNA binding	2.5E-06	3.2E-08

ND Not determined

the post-synaptic density, critical to the transmission of nerve impulses (Table 1).

Herein, we report the analysis of the distribution of tetranucleotide repeat (TR) sequences ≥8 units [89] in human genes. Of the 29,708 TR tracts found genome-wide [89], 8,943 (∼1/3) were located in 4,182 non-redundant RefSeq genes (∼1/5 of all annotated genes), or within 1 kb of their transcriptional boundaries, with an average of 2 TR tracts per gene. Also, 114 genes were found to contain the repeats in the promoter region (within 1 kb of the predominant transcription start site), 2 in the 5′-UTR, 4,485 in introns, 23 in the 3′-UTR, and 100 within 1 kb downstream of the transcriptional unit. Thus, ∼95% of gene-associated TRs are located within introns. The group of TR-containing genes was found to be most enriched for genes involved in cell adhesion, localization to the plasma membrane, ion channel function, and receptors involved in signal transduction pathways, cell communication, and transmission of the nerve impulse (Table 1 and Electronic Supplementary Material, ESM). In addition, genes associated with glutamate receptor activity were progressively enriched as a function of TR length (ESM Fig. 1 and ESM text).

The enrichment analyses for the two gene datasets containing either TR (≥8 units, 4,182 genes) sequences or long R:Y tracts (≥100 nt, 1,951 genes) were extended to additional genomic functions [93]. Both datasets were highly enriched in genes known to undergo alternative splicing and prone to DNA breakage leading to chromosomal translocations (Table 1). These data enable the following conclusions: (1) the categories of genes enriched in long R:Y tracts are also enriched in TR sequences; (2) the gene functions involved are associated, as a whole, with communication between cells; (3) long R:Y tracts (which also include most, TRs ≥18 units, ESM and ESM Fig. 1) are an exquisite property of synaptic glutamatergic activity; (4) intragenic R:Y and TR tracts are characteristic of genes that have acquired a complex organization through alternative splicing and, thus, may encode proteins with multiple functions, and (5) the genes involved are generally prone to breakage. An important aspect of these studies is the association between R:Y-tract containing genes and genes that confer susceptibility to complex mental disorders [74]. This association has recently been strengthened by genome-wide case–control analyses [94] in subjects afflicted with schizophrenia [74, 95]. Hence, triplex-forming sequences are attributes of genes involved in integrative networking functions in the brain.

Analysis of the distribution of micro- and minisatellites ranging from dinucleotides to 11-mer repeats in human cDNAs [89] identified 2,626 unique RefSeq genes. The set displayed strong enrichment for genes associated with transcription factors, the regulation of transcription and specific signaling pathways, including genes from the MAPK and WNT pathways (Table 1). Similar searches at the proteomic level also showed preferential enrichment for transcription factors, chromatin binding proteins, DNA and RNA binding proteins, and proteins involved in translation [96, 97]. The current rationale for these observations consists of a model whereby homo-amino acid runs constitute disordered protein regions that become ordered upon nucleic acid and/or cognate protein binding. The transition from a disordered to an ordered state would then greatly enhance the stability of the ensuing complexes and therefore elicit specific biological functions [29].

As mentioned above, G-quadruplex-forming repeats predominate in gene regions flanking the transcription start sites but are also abundant in 3′-UTRs. The classes of genes most enriched in such repeats belong to the family of small GTPases, such as Rho, which play critical roles in signal transduction [98] and in the regulation of stress fibers, including the actin cytoskeleton [99] (Table 1).

In summary, the association of repetitive DNA sequence with gene function follows specific patterns, i.e., genes involved in male reproduction for large IRs, cell–cell communication for long R:Y tracts, transcription and its regulation for coding microsatellites and small GTPase signaling/regulation for G-quadruplexes. Therefore, it is likely that selective pressures have acted so as to maintain specific DNA sequences in coding regions to enable the acquisition and maintenance of novel gene functions during the course of evolution.

Patterns of global gene expression

The first analyses on the genome-wide distributions of quadruplex-forming motifs ($G_{3+}N_{1-7}$ $G_{3+}N_{1-7}$ $G_{3+}N_{1-7}$ G_{3+}) revealed their high prevalence in warm-blooded species [100] and an overrepresentation in the promoter region of genes [75, 80, 82, 101]. Indeed, a recent investigation on a dataset of 13,276 non-redundant human RefSeq genes established the presence of one or more G4 motifs in the 500-nt region flanking the transcription start site (TSS) of 8,214 (∼62%) such genes [81], a significant proportion. When the expression value of the RefSeq genes was analyzed in 79 human tissues/cell types, a significant association was found between G4 motifs downstream, but not upstream, of the TSS and an increase in gene expression. Moreover, a direct relationship was evident between the number of G4 motifs (0–4) and the levels of gene expression. Further analyses indicated that the average levels of gene expression for both the G4-negative and G4-positive genes varied according to tissue/cell type. Nevertheless, in each case, the G4-positive gene set displayed higher transcriptional values than the G4-negative set (Fig. 2a). Hence, a direct association exists between G4 motifs and gene transcription, supporting a genome-wide

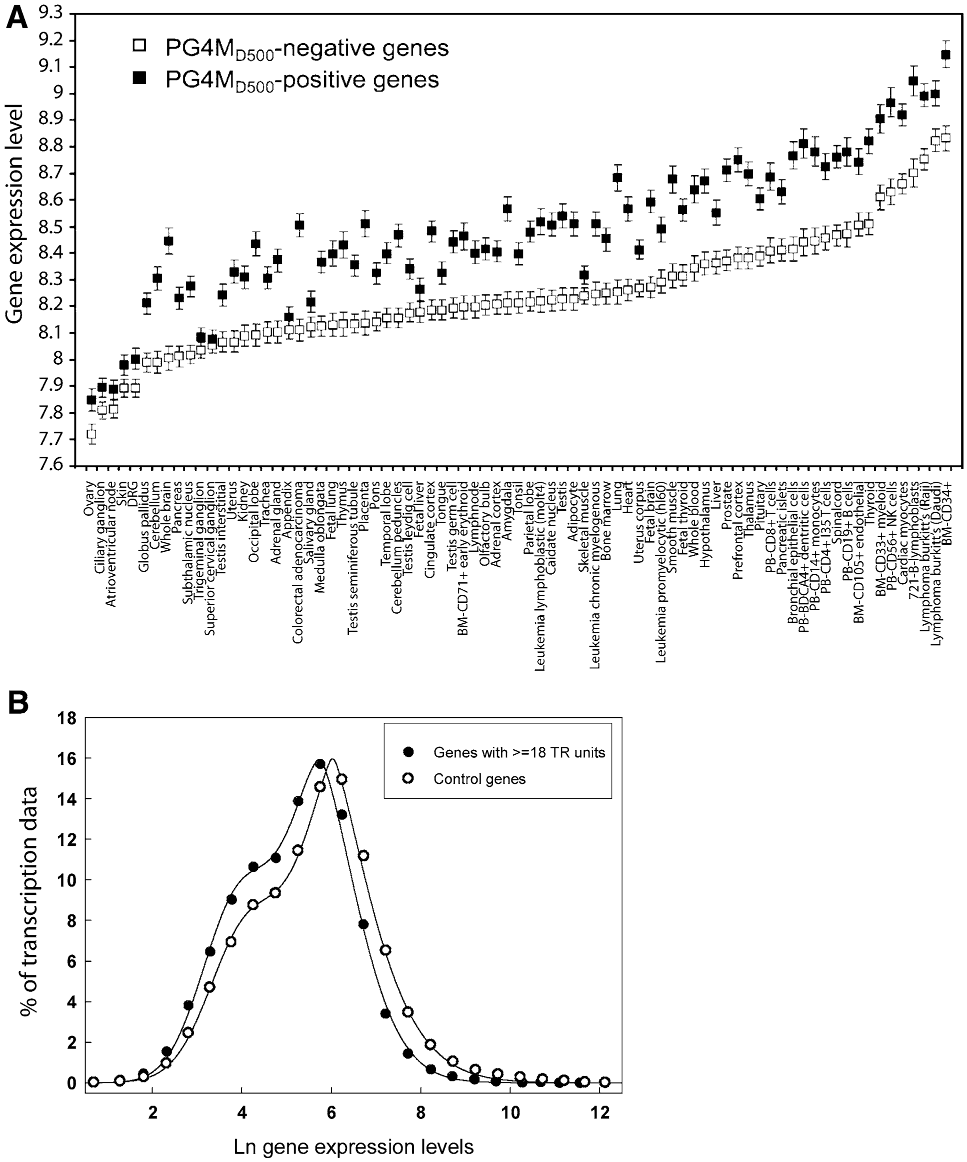

Fig. 2 **a** Expression profiles of quadruplex-containing genes. Comparison of the gene expression levels between genes containing quadruplex-forming sequences (PG4M$_{D500}$-positive, *filled squares*) and genes without PG4M$_{D500}$ (*open squares*) in each human tissue/cell type. *Error bars* represent the 95% confidence interval of the mean expression level. (Reprinted with permission from [81].) **b** Global gene expression profiles of genes containing triplex-forming sequences. Each data point represents the mean ln (*x*-axis) for all gene expression values falling within 0.5 ln-interval bins, from 0–0.5 to 12.0–12.5. On the *y*-axis is the percentage of the gene expression values falling within each 0.5 ln-interval bin relative to the total number of gene expression values for either the control genes (*open symbols*) or the genes harboring ≥18 TR units (set 2) (*filled symbols*)

role for quadruplex structures in either promoting transcriptional activity and/or stabilizing the ensuing pre-mRNA transcripts.

Quadruplex nucleic acid structures are likely to regulate transcriptional activity by several, and perhaps opposing, mechanisms. A recent search for G4 motifs in 32,985 annotated 5′-UTRs and 32,818 3′-UTRs from a compilation of 21,658 human genes yielded the following trend in relative frequencies per kb of DNA: 5′-UTR > 3′-UTR > transcriptome > whole-genome, with values ranging from 0.382 to 0.057 [83]. Significantly, not only G4 motifs were overrepresented in the 3′-UTRs in addition to 5′-UTRs, but also for a high proportion of genes (97/561 or ~17%) with G4 motifs in 3′-UTRs, the genomic distance from the end of transcription to the next gene was shorter (within 1 kb) than genome-average, suggesting a role for G-quadruplex structures in transcription termination. Finally, a large body of evidence [102] supports the conclusion that quadruplex DNA may form in the promoter region of oncogenes and elicit functional roles, such as the transcriptional inhibitory activity observed in *c-MYC* [103, 104].

Herein, we contrast the global gene expression profile of genes that contain quadruplex-forming sequences with those that harbor triplex-forming sequences, i.e., the set of 228 genes (set 1) containing the longest (≥250 nt) R:Y tracts (Table 1) and the set of 190 genes (set 2) containing ≥18 TR units (Table 1; ESM and ESM Fig. 1). Analysis of the gene expression data in 70 tissues/cell lines (cancer tissues and cancer cell lines were not included) showed that for the 16,146-probe set comprising the control genes (i.e., sets 1 and 2 excluded) the transcriptional values followed a bimodal distribution composed of two overlapping Gaussian curves (ESM Fig. 2), the first accounting for 75% of the data and showing high levels of gene expression (HGE) and the second comprising the remaining 25% of the data and displaying low levels of gene expression (LGE) (Fig. 2b). For comparative purposes, the HGE mean value was normalized to 1. Accordingly, the LGE mean value was 0.13 when the respective natural logarithms were transformed in raw gene expression data, a ninefold reduction. Set 1 also displayed a bimodal distribution. However, whereas the LGE mean value did not differ from the control probe-set, the HGE distribution was shifted to significantly lower values (normalized mean = 0.73, $P < 0.001$, a $\sim$25% reduction relative to the control data-set mean). Similarly, set 2 displayed significant reduction in gene expression for both the HGE and the LGE distributions (normalized mean values 0.75 and 0.11, respectively; $P < 0.001$) (Fig. 2b). Hence, genes containing long R:Y tracts with the potential to form triplex DNA structures are generally transcribed at lower levels than genes that do not contain such elements. A previous analysis [74] of the tissue-specific patterns of gene expression for set 1 after z-scoring (which normalizes the average expression of any given gene across all tissues) indicated that the highest transcriptional activity occurred in the brain. Hence, taken together, these data suggest brain-specific roles for long R:Y tracts in transcriptional regulation. Finally, these analyses reveal the contrasting transcriptional profiles of genes harboring quadruplex-forming repeats (increased transcription) and those containing triplex-forming sequences (decreased transcription).

Cruciforms and the genomic architecture of the human Y-chromosome

Sex-specific genes are clustered in the arms of IRs on the X and Y chromosomes [92]. The Y-chromosome comprises two external pseudo-autosomal (PAR1 and PAR2) regions ($\leq$1.5 Mb) homologous to the X-chromosome and essential for chromosome segregation at meiosis, and a central male-specific segment (MSY) functionally divided into euchromatin (shorter p-arm) and heterochromatin (distal q-arm).

The euchromatin region is itself a complex mosaic of modular DNA sequences characterized by eight large (up to 1.46 Mb in length) inverted repeats, commonly referred to as palindromes 1–8, shorter inverted and direct repeats, all of which contain gene families with expression patterns specific to the testis and performing essential functions in the production and maturation of sperm (Table 1 and [92]). Two other regions, X-transposed and X-degenerate, harbor paralogous genes with copies on the X-chromosome. Modular tandem arrays also compose the entire heterochromatic region, whose length variation caused by polymorphic tandem array repeat number confers large-scale differences to the size of Y-chromosomes in the general population (Fig. 3a). Hence, inverted and direct repeats comprise most of the human Y-chromosome, thus conferring higher-order structural architectures to the primary genomic sequence.

The MSY region does not have a counterpart in other chromosomes and thus it is excluded from sexual recombination. This unique behavior has prompted speculation [105] that Y-chromosome extinction is inevitable given that gene decay, consequent to naturally occurring mutations, would be irreversible. Indeed, the Y-chromosome has degenerated substantially both in size and gene content in comparison with the X chromosome. However, the ampliconic gene families nested within the palindromic arms and key to spermatogenesis have sustained much lower than expected mutation rates during evolutionary time [106]. For example, not only the intrapalindromic (arm-to-arm) sequences share on average >99% sequence identity, but also gene pairs located at symmetrical positions within palindromic arms are generally identical or nearly so [107]. In contrast, substantial sequence divergence exists between gene pairs belonging to the same gene families but located at different arm positions [107]. Thus, high rates of gene conversion are believed to have occurred among testis-specific genes in the human Y-chromosome, which have effectively counteracted the threat of gene decay imposed by the absence of meiotic recombination [106, 108]. In fact, comparative analyses between the human and chimpanzee Y-chromosomes strongly support the conclusion that the ampliconic gene families in palindromes have been under strong positive selective pressure, most likely because of their key role in spermatogenesis (Fig. 3a) [92].

These observations raise a number of questions. Did the inverted repeat architecture of palindromes play a critical role in shaping and preserving Y-chromosome function? How did gene conversion take place between the arms of palindromes? Several studies have been performed to address these issues. First, analyses from representative ethnic groups revealed that the IR3/IR3 region of the Y-chromosome was inverted in 16/47 cases [92]. This corresponds to a frequency of $\sim 9.2 \times 10^{-4}$ inversion events per father-to-son transmission, a frequency that is at least 10,000 times higher than that of single nt changes. Second, recent detailed sequence analyses of microinversions that distinguish the human and chimpanzee genomes showed that in all cases inverted repeats were present at breakpoints [109]. Therefore, whereas inverted repeats may suppress random nucleotide changes arising from within their repeating arms [107], they nevertheless represent a structural unit capable of changing genomic orientation over time. We and others [110] have proposed that large inverted repeats may promote strand exchange and form stem-loop structures, which may account for these features

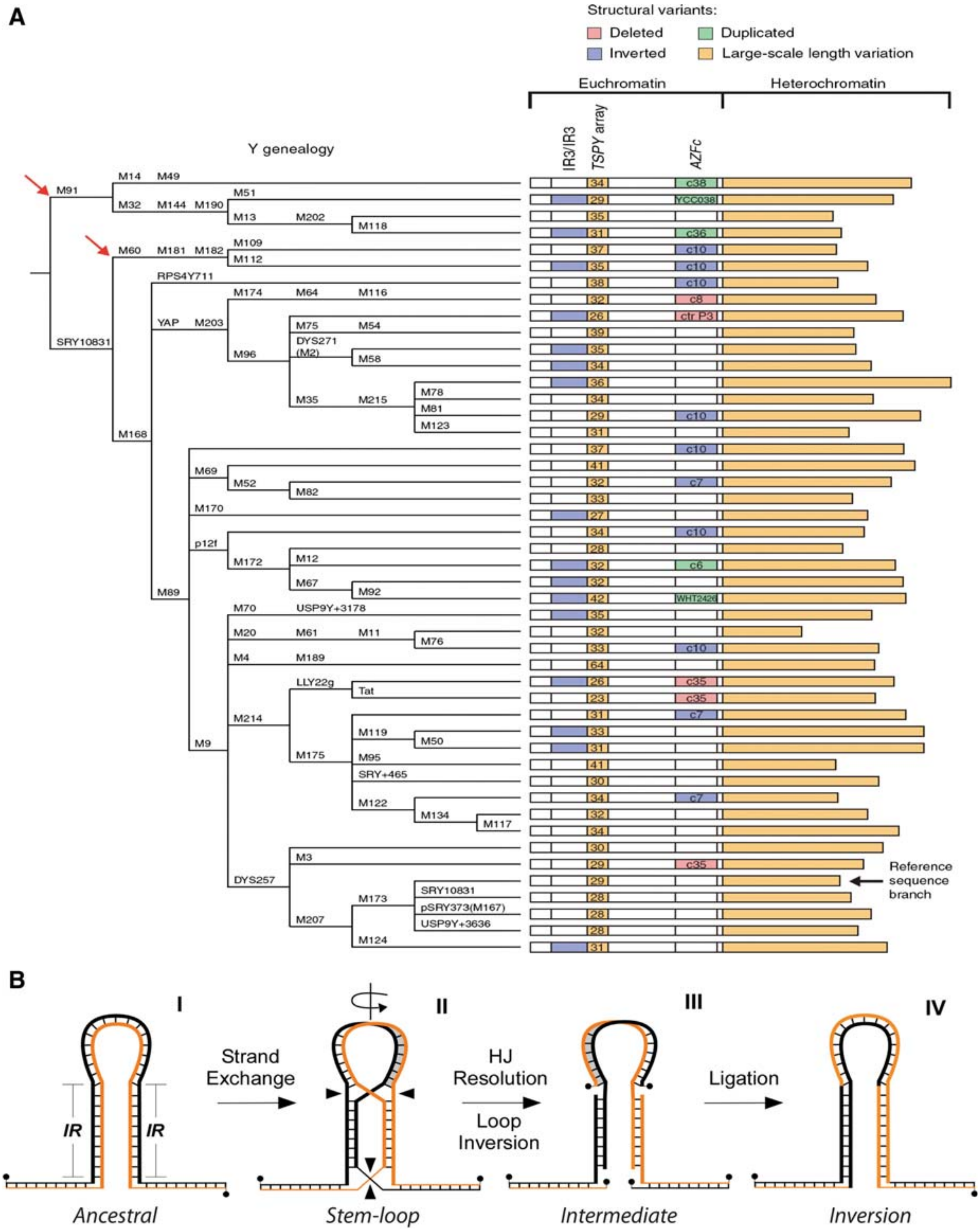

Fig. 3 a Y chromosome genealogical tree (*left*) and identified structural polymorphisms (*right*). Chromosomes were assigned to one of 47 branches by typing for the stable, biallelic polymorphisms indicated. *Red arrows* indicate major branches confined to Africa. For each branch, the structure of the Y chromosome sampled is schematized, including (*far right*) the length of distal-Yq heterochromatin. Within the euchromatin, the presence of a particular structural variant is indicated by a *color-coded rectangle*. (Reprinted with permission from [92].) **b** Model for stem-loop-mediated chromosomal inversion and strand exchange. *Structure I* illustrates the original (*ancestral*) sequence organization with two inverted repeat (*IR*) segments. *Structure II* shows the stem-loop structure containing two Holliday-like junctions originating from strand exchange and the inverting loop. *Structure III* represents the intermediate DNA species after Holliday junction (HJ) resolution and loop inversion. *Structure IV* depicts the final DNA configuration with the complementary strands containing DNA bases located originally on the same strand and the inverted loop. (Adapted from [109])

[109]. Accordingly (Fig. 3b), the two arms of an inverted repeat may interact and engage in a strand-exchange reaction leading to the formation of intra-strand Watson–Crick hydrogen bonded base pairs (Fig. 3b, Structure I). This gives rise to a stem-loop structure characterized by two Holliday-like junctions, one at the apex between the stem and the looped-out intervening sequence, the other at the base between the stem and the sequences flanking the inverted repeats (Fig. 3b, Structure II). Resolution of the Holliday-like junctions would yield two types of events. First, in 50% of cases, the intervening sequence will invert, assuming equal rates of cleavage at the intersecting versus non-intersecting strands (Fig. 3b, Structure III). Second, upon inversion, the DNA complementary strands of the inverted repeats will contain the nucleotides that were previously located on the same DNA strand, effectively providing a means for the correction of mispairs, through mismatch or other repair pathways (Fig. 3b, Structure IV). These models (Fig 3b and [110]) offer a rationale for the observations that: (1) inverted repeats mediate genomic inversions [109]; (2) high rates of "gene conversion" events take place between the arms of palindromes [106]; and (3) genes of the same family show a pair-wise pattern of sequence identity based upon their location at similar palindromic arm position [107]. In addition, these structures provide a model for the formation of large stem-loop structures, including cruciforms [24–28, 111], for which the physiologic levels of negative supercoiling appear insufficient [112]. Finally, because strand exchange may initiate and terminate anywhere along the inverted repeat sequences, their total lengths do not impose a size constrain to stem-loop structures, which may vary in length. This contrasts with the "classic" cruciform structure (Fig. 1a), which nucleates from the apical loop.

In summary, these composite data provide empirical evidence in support of the notion that cruciforms have played a pivotal role during evolutionary time by providing a genomic structure upon which selection acted so as to preserve, and perhaps shape, the sex-specific functions of the human Y-chromosome.

Mechanisms of DNA structure-induced genomic instability

Studies using model systems suggest that instability caused by trinucleotide repeats and other non-B DNA-forming sequences may occur via aberrant DNA replication events [16, 113, 114], as well as replication-independent mechanisms in non-proliferating tissues [115]. We discuss results to support both replication-dependent and replication-independent mechanisms of DNA structure-induced genetic instability below.

Replication

Human fragile sites often consist of non-B DNA-forming tandem repeats [22]. Studies of model sequences have provided links between DNA replication and fragile site instability [114, 116]. For example, the mutation rate of hairpin-forming CAG repeats increased when the DNA polymerase zeta subunit rev1 was mutated in *Saccharomyces cerevisiae* [117], suggesting that the transient formation of single-strand DNA during replication and the ensuing slipped DNA structures are mutagenic. Indeed, replication slippage at repetitive sequences (e.g., CTG:CAG, GAA:TTC, CGG:CCG, and GAC:GTC) has been implicated in mutations, deletions, or expansions of repeating units, causing genetic instability related to hereditary neurological diseases [15].

Replication stalling

Direct evidence for a link between replication and non-B DNA structures was provided by the ability of non-B DNA structure-forming sequences to slow replication forks. Using two-dimensional gel electrophoretic analyses and electron microscopy, stalling of replication intermediates by trinucleotide repeats, inverted repeats of *Alu* elements [118], and an (A + T)-rich fragile site (FLEX1) from the human *FRA16D* gene [119] was detected when these elements were cloned into bacterial, yeast, and human cells. Replication attenuation was dependent on the length and/or sequence of these repeats and correlated with their capacity to form DNA secondary structures. A stalled replication fork will give rise to longer exposure of single-stranded DNA, and may cause replication fork collapse and DSBs, which may be processed in a mutagenic fashion. DNA triplex structures can also block replication forks and cause DSBs [12, 42].

Orientation of repeat sequences

Due to the differences between leading and lagging strand DNA synthesis during replication, the orientation of repeat sequences greatly influences their stability in model systems such as bacteria, yeast, and cultured mammalian cells [120–123]. Most non-B DNA structure-forming trinucleotide repeats are more unstable when they serve as lagging strand templates. The instability of GAA repeats in the *FRDA* gene responsible for Friedreich ataxia is dependent on the orientation of DNA replication. In yeast, for example, GAA repeats display nearly 100-fold higher instability on the lagging strand than on the leading strand [124]. Similarly, CTG repeats show higher levels of DNA instability when used as a template for lagging strand synthesis (to the replication origin ColE1) in a *recA*$^{-}$ strain of *E. coli*, upon induction of DSBs [125]. A long $(CTG)_{130}$ repeat from a myotonic dystrophy patient was unstable on the lagging-strand template but was stable on the leading strand template in yeast [123]. Also, the $(CGG)_{160}$ repeat from the 5′-UTR region of the *FMR1* gene contracts when placed as the lagging strand template in the yeast chromosome, but yields few contractions when the repeat is located in the leading strand template [120]. The strand-preference of trinucleotide repeat instability indicated that the ability to form secondary structures differs for the two complementary sequences. For example, the CTG repeats adopt a more stable hairpin structure than CAG repeats [126, 127]. Hence, when CAG repeats serve as the lagging strand template, the newly synthesized complementary CTG repeats would be prone to form non-B structures that may cause repeated synthesis, resulting in expansion of the repeat [15, 17]. At the same time, if the leading strand template with CTG repeats forms secondary structures, it may be bypassed and give rise to contractions within the repeat. Whereas, contractions of trinucleotide repeats are seen in many yeast and bacterial models, expansions are prevalent in human diseases [14, 15, 128]. The reasons for this discrepancy remain to be clarified; however, transacting factors may be involved. For example, the human MSH2–MSH3 complex can bind CAG or CTG repeats [129], and knockdown of the proteins in this complex has been shown to reduce trinucleotide repeat instability [130, 131]. Thus, it is possible that the MSH2–MSH3 complex might stabilize the repeats rather than processing the “mismatched” nucleotides. Due to its strand discrimination ability, MSH2–MSH3 might then stabilize the structure formed on trinucleotide repeat tracts on the newly synthesized strand preferentially, leading to expansion events.

Replication proteins

The ability of non-B DNA structure-forming sequences to stall replication forks can be counteracted by proteins that stabilize replication forks. Studies on CGG repeats and inverted repeats in yeast indicate that the replication fork-stabilizing proteins Mrc1 and Tof1 could reduce the replication stalling effect of non-B DNA structures [118, 132]. Proteins functioning in the maturation of Okazaki fragments also influence the expansion and contraction of repeat sequences. For example, mutations in yeast Rad27 (homologous to the human FEN-1 flap endonuclease 1) lead to the expansion of repeated CAG:CTG sequences and to the recombination/instability of inverted *Alu* elements [133, 134]. The interactions among Rad27, DNA ligase I, and proliferating cell nuclear antigen (PCNA) are critical for the maintenance of CAG:CTG repeats in yeast [135]. Similar to Rad27, which prevents the expansion of trinucleotide repeats, the yeast helicase Srs2 unwinds the secondary structures formed by trinucleotide repeats and, together with post-replication repair proteins, prevents the expansion of CAG:CTG repeats [136–138]. However, these results demonstrating a role for Rad27 in repeat stability in yeast are not consistent with those observed in mammalian cells. For example, the CAG:CTG repeat from the Huntington locus was stable over 27 successive cell passages when *FEN-1* was continuously knocked-down by siRNA [139]. Similarly, in mice, haploinsufficiency of Fen1 increases the expansion of CAG:CTG repeats at the Huntington’s locus but does not affect their stability at the myotonic dystrophy type 1 (DM1) locus in knock-in models [140].

Whereas DNA replication-related mechanisms may largely be responsible for non-B DNA structure-induced genomic instability in proliferating tissues, they do not account for genetic instabilities found in non-proliferative tissues [115]. For example, analyses of patients with Huntington disease and spinocerebellar ataxias showed instability of CAG:CTG repeats in their non-proliferative tissues, such as brain and sperm [141, 142]. Similarly, H- and Z-DNA structures were found to induce large-scale deletions and rearrangements in replication-deficient HeLa cells ([18], and our unpublished results). In transgenic mice CAG repeats might expand by gap repair in germ cells without replication or recombination taking place [128, 143]. In addition, the translocation of the palindromic AT-rich repeat has been shown to be independent of replication [25, 111]. Several DNA repair-related mechanisms have been proposed to explain replication-independent mutagenesis events at non-B DNA conformations [115].

Recognition of non-B DNA structures

Being different from the canonical B-form DNA conformation, non-B DNA structures represent distortions of the DNA double helix, including the non-B structure itself, and the non-B to B-form junctions. These distortions may be recognized as “damage” by DNA repair proteins. One consequence of such “damage” recognition is the introduction of mutations/deletions, causing genomic instability (Fig. 4). Many non-B DNA structures can lead to the generation of DSBs during DNA repair, which are critical lesions that can lead to cell death or chromosomal rearrangements [10].

Hairpins/cruciforms

Trinucleotide repeats can form hairpins with mismatched nucleotides in the stems. This structural property may be recognized as “damage” by repair proteins. The Mre11/Rad50 complex was shown to cleave hairpins/cruciforms in

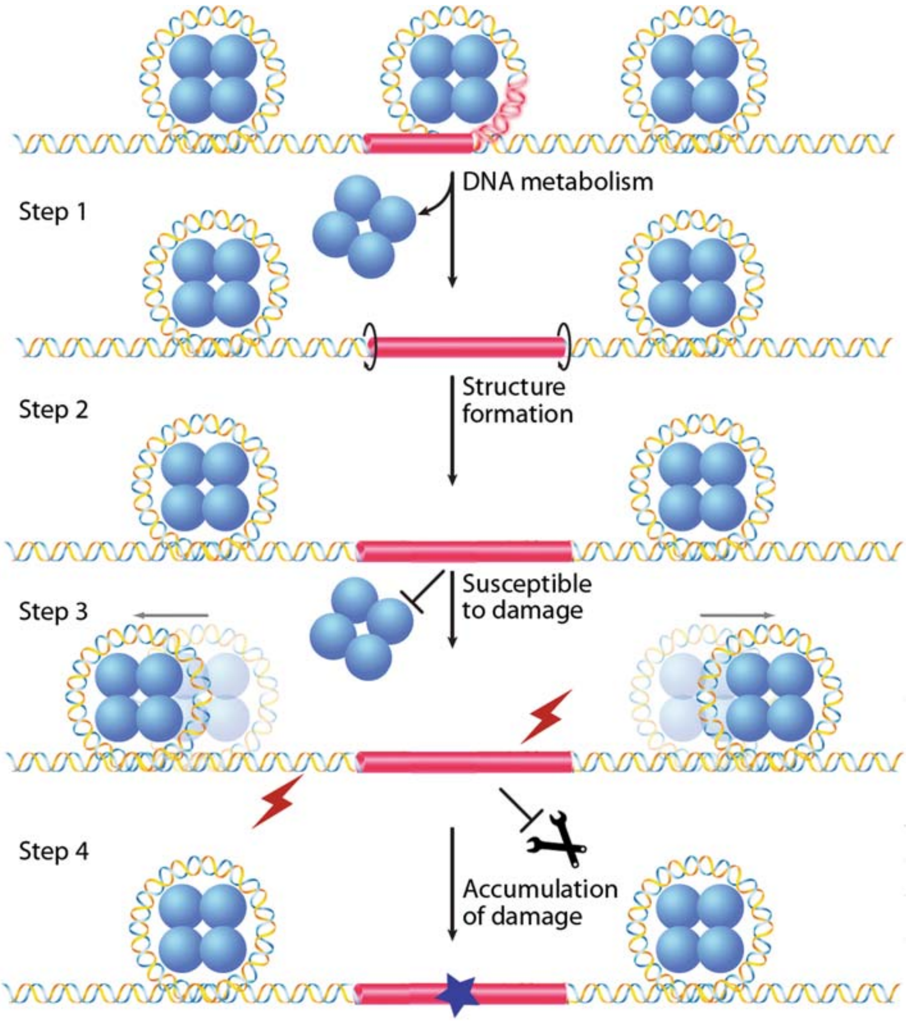

Fig. 4 DNA damage and non-B DNA structures. Unwrapping of a non-B DNA-forming sequence (*red box*) from the histone core during DNA metabolism (*Step 1*), facilitates the non-B DNA conformation (*Step 2*). The non-B DNA conformation may be more susceptible to DNA damage and the damage in the non-B DNA region may be more resistant to repair (*Step 3*), leading to accumulated damage (*blue star*) in this region (*Step 4*). (Adapted from [115])

a structure-specific manner [144]. Inverted repeats also generate DSBs and stimulate unequal sister-chromatid exchange in yeast [129]. Although it was not evaluated whether replication is important for DNA breakage and translocation in this case, the rate of this spontaneous exchange was reduced to ~50% in yeast strains with mutations in the mismatch repair (MMR) genes *Msh2* or *Msh3*, suggesting a role for DNA repair in non-B DNA structure-induced mutagenesis [145]. Kirkpatrick and Petes [146] reported that repair of 26-base loops in yeast involved both Msh2 and Rad1, suggesting that these repair proteins recognize helical distortions and remove DNA loops formed by trinucleotide repeats. The absence of a functional nucleotide excision repair (NER) protein UvrA has been shown to increase the instability of long CTG repeats in *E. coli* [146, 147]. However, conflicting results on the roles of MMR and NER repair proteins on repeat instability have been reported in human cell lines or mouse model systems [130, 148–150]. Thus, further studies in this area are warranted.

Z-DNA

While it is clear that Z-DNA-forming sequences can cause genetic instability in a number of organisms, the underlying mechanisms remain largely speculative [151]. Studies from our laboratory have demonstrated that the instability of Z-DNA-forming sequence $(CG)_{14}$ results from the DSBs induced by these sequences in mammalian cells [18]. However, the mutation spectrum induced by the same $(CG)_{14}$ sequence in bacteria is quite different [18]. In bacteria, the predominant mutation/deletion appears to be within the CG repeat with a gain or loss of dinucleotides, likely caused by slippage events during replication. In contrast, replication was not required for the $(CG)_{14}$-induced mutations in mammalian cells, where predominant mutation events were large (>50 bp) deletions [18]. It is possible that these deletions were the result of error-generating DNA repair processing events at these unusual DNA structures. Chromatin immunoprecipitation experiments showed that Z-DNA-forming $(CG)_{14}$ repeats were enriched relative to B DNA sequence controls in the precipitations with antibodies against the NER protein, XPA, and the MMR protein, MSH2 (unpublished data). Moreover, the mutation frequency of this Z-DNA-forming sequence was lower in XPA- or MSH2-deficient human cells than in their isogenic wild-type counterparts, suggesting that these proteins contribute to Z-DNA induced mutagenesis in human cells (unpublished data).

H-DNA

We have demonstrated that the naturally occurring H-DNA structure-forming sequence from the human *c-MYC* gene, which co-localizes with translocation breakpoints, can induce DSBs within these sequences in mammalian cells and cause genomic instability in mice [12, 23]. The instability of H-DNA structure-forming sequences from the polycystic kidney disease 1 (*PKD1*) gene was lower in MMR-deficient bacterial cells compared to wild type cells [10]. Our data suggest that like Z-DNA, the mutagenicity of H-DNA-forming sequences involves XPA and MSH2 (Wang and Vasquez, unpublished data). Recently, we discovered that the MMR protein complex, MSH2–MSH3 (MutSβ), cooperates with two key NER protein complexes (XPA-RPA and XPC-RAD23B) in the recognition of triplex structures in the presence of a psoralen interstrand crosslink. This interaction was enhanced up to tenfold in the presence of a psoralen interstrand crosslink within a triplex structure compared to a psoralen interstrand crosslink within a duplex DNA substrate, suggesting that the non-B DNA structure is a strong recognition signal for both NER and MMR proteins [152].

However, binding of DNA repair proteins to non-B DNA structure-forming sequences does not always result in increased instability. In some cases, binding of MSH2 or MSH3 to the hairpin structures formed by trinucleotide repeats may prevent the structure from being processed. In yeast, the Msh2–Msh3 complex binds preferentially to the imperfect stem formed by interrupted trinucleotide repeats and blocks their expansion [153]. The human MMR protein complex MSH2–MSH3 was confirmed to preferentially bind looped-out secondary structures formed by CTG repeats, and the ATPase activity required for its repair function was decreased after binding to the non-B DNA structure-forming sequences [129].

DNA repair and non-B DNA structure-forming sequences

DNA repair processes may promote the transition from B- to non-B DNA structures. When DNA damage occurs at or near repeated sequences, the subsequent repair processes may unwrap the DNA from the chromatin, which generates negative superhelical stress and promotes the transition to non-B DNA. Alternatively, single-stranded DNA regions may form, which then allow the folding of secondary structures to take place (Fig. 4). Genetic experiments in a mouse system demonstrated that knockdown of the recombination protein Rad52 decreased the expansion of CTG repeats [154]. Introducing DSBs within the GAA repeats or within CTG repeats in *E. coli* resulted in deletion, but this stimulatory effect only occurred when DSBs were located within the repeats [155, 156]. Similarly, more instability was seen in the processing of DSBs with a CTG repeat sequence in mammalian cells when the CTG repeat was capable of forming slipped DNA structures compared to a linear DNA control [157]. These results suggest that hairpin/cruciform structure-forming sequences may be more susceptible to deletion or rearrangement events during DNA repair in the surrounding regions.

On the other hand, the formation of DNA secondary structures near DNA damage might influence the repair processing, depending on the type of damage, the environment, and the nature of the secondary structures. For example, the Malkova [158, 159] group has shown that, in yeast, the inverted Ty elements promote the repair of DSBs at distances of up to 30 kb from the elements by forming dicentric inverted dimers. The existence of inverted repeats flanking a DSB is thought to channel repair from a homologous recombination pathway into a single-strand annealing-gross chromosomal rearrangements (SSA-GCR) pathway in yeast [158]. This pathway is not dependent on homologous recombination because, in a *rad51Δ* strain, the existence of intact large inverted repeats near the DSB reduced the broken chromosomal loss from roughly 40 to ~13% [158]. Unlike inverted repeats which promote the repair of DSBs, the secondary structures formed by CTG units in a plasmid reporter system in mammalian cells showed decreased repair efficiency of the DSB within the repeat, compared to a control of linearized plasmid containing the same CTG sequence and DSB [157]. These results suggest that non-B DNA structures are able to form during DNA repair and that the formation of such structures can potentially alter repair. If the non-B DNA structure-forming sequences near the damage site are processed during the repair of the lesion, they may contribute to the error-generating repair and lead to genomic instability. This notion is supported by data from patients showing that gene conversion contributes to the instability of CGG:CCG repeats in the *FRAXA* and CTG:CAG tracts in DM1 cases [160].

Non-B DNA structures may also affect DNA repair by increasing DNA damage susceptibility and/or damage accumulation [115]. The distortion of the DNA helix and the altered arrangement of the bases and sugar moiety in non-B DNA conformations can influence the interactions of DNA damaging factors with the nucleotides, and thus modify their accessibility to DNA damage. For example, many types of non-B DNA conformations, e.g., H-DNA, B–Z junctions, and hairpin/loop structures, contain single-stranded regions that are not protected by hydrogen bonding and are often precluded from chromatin that can otherwise protect the bases. Thus, non-B DNA structures may be more accessible to DNA damaging factors than B-DNA [115]. For example, the guanines in a Z-DNA structure are more sensitive to ionizing radiation [161], and are more sensitive to oxidative damage in the single-stranded regions compared to B-form duplex DNA [162]. On the other hand, it is also possible that DNA in non-B conformations are more resistant to certain types of damaging agents, e.g., interstrand crosslinks are less likely to be formed in the single-stranded regions of non-B-DNA structures than in duplex DNA.

The abnormal positioning of the bases and sugar moiety in non-B-DNA conformations can also impact the function of some DNA repair proteins on damaged DNA. For example, alkylating damage such as *N*7-methylguanine or *O*6-methylguanine is not repaired as efficiently in Z-DNA as it is in B-DNA [163, 164]. This topic is covered in depth in a recent review by Wang and Vasquez [115], which describes a model of “DNA repair-stimulated non-B DNA structure formation”.

Concluding remarks

Since the discovery of non-B DNA structures several decades ago, these structures have been shown to influence critical genetic transactions, such as DNA replication, transcription, recombination, and repair. Our knowledge of the role of non-B DNA structures in genomic instability has recently been gained along with the progress made in understanding the DNA structural characteristics, the correlations between DNA structure and genetic diseases, and the proteins that influence the stability of DNA structures. Genome-wide analyses have greatly influenced our view on DNA structure-induced genomic plasticity and its consequence in human disease and on the evolutionary changes since the divergence from prokaryotes to eukaryotes. The capability of non-B DNA structures to induce mutations/deletions and to promote chromosome rearrangements gives them potential evolutionary functions; e.g., mutating to adapt to rapid changes and at the same time, keeping DNA information through recombination (in the case of the human Y chromosome mentioned above).

However, there are still many questions to be answered regarding the relationships between DNA sequence, structure, and function. For example, what environmental conditions promote non-B structure formation? What proteins function in the recognition and subsequent processing of non-B DNA structures? What proteins/pathways are involved in their error-generating repair causing genomic instability? The same trinucleotide repeat sequences in various systems do not always result in genetic instability, suggesting that DNA sequence context and/or location in the genome may be critical factors in repeat instability. In our studies, H-DNA sequences are mutagenic in mammalian cells, but are not mutagenic when introduced in bacteria, suggesting a requirement for transacting factors/proteins in a host-specific fashion for structure formation and/or processing. The observation that specific types of non-B DNA-forming sequences are enriched in gene families with particular functions, and the correlation between gene expression levels, and the presence of non-B DNA-forming sequences in these gene regions, emphasizes the need to further investigate the regulatory function of repetitive elements. It is not clear whether these elements are enriched due to their regulatory function or due to the higher mobility of unstable non-B DNA structure-forming sequences.

The current mechanisms proposed for non-B DNA-induced genetic instability include abnormal DNA replication that can explain the contraction and expansion of trinucleotide repeats in replicating systems and the processing by DNA repair proteins that contribute to replication-independent mutagenesis induced by non-B DNA structures. Many DNA repair proteins have been found to interact with non-B DNA structures in vitro; while some protein–non-B DNA interactions lead to repair processing and DNA breakage, other proteins might stabilize the non-B DNA conformations. Furthermore, a particular protein may have different affects on non-B DNA conformations in different species. The much-needed screening for proteins that interact with non-B DNA structures is in progress and will provide more information about their recognition and structure-induced genomic instability at the molecular level. These results will help us to comprehensively understand how these DNA structures influence genome stability, DNA metabolic functions (e.g., gene function and regulation), and the balance between selection stress and adaptation to changing environmental conditions.

Acknowledgments We thank Ms. Sarah Henninger for technical assistance. This work was supported by an NIH/NCI grant (CA093729) and an NIH/NIEHS grant (ES015707).

References

1. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737–738

2. Mirkin SM (2008) Discovery of alternative DNA structures: a heroic decade (1979–1989). Front Biosci 13:1064–1071
3. Felsenfeld G, Davies DR, Rich A (1957) Formation of a three-stranded polynucleotide molecule. J Am Chem Soc 79:2023–2024
4. Wang AH, Quigley GJ, Kolpak FJ, Crawford JL, van Boom JH, van der Marel G, Rich A (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. Nature 282:680–686
5. Lilley DM (1980) The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. Proc Natl Acad Sci USA 77:6468–6472
6. Panayotatos N, Wells RD (1981) Cruciform structures in supercoiled DNA. Nature 289:466–470
7. Lyamichev VI, Panyutin IG, Frank-Kamenetskii MD (1983) Evidence of cruciform structures in superhelical DNA provided by two-dimensional gel electrophoresis. FEBS Lett 153:298–302
8. 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
9. Bacolla A, Wells RD (2004) Non-B DNA conformations, genomic rearrangements, and human disease. J Biol Chem 279:47411–47414
10. Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova N, Abeysinghe SS, O'Connell CD, Cooper DN, Wells RD (2004) Breakpoints of gross deletions coincide with non-B DNA conformations. Proc Natl Acad Sci USA 101:14162–14167
11. Wang G, Vasquez KM (2006) Non-B DNA structure-induced genetic instability. Mutat Res 598:103–119
12. Wang G, Vasquez KM (2004) Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. Proc Natl Acad Sci USA 101:13448–13453
13. Glickman BW, Ripley LS (1984) Structural intermediates of deletion mutagenesis: a role for palindromic DNA. Proc Natl Acad Sci USA 81:512–516
14. Orr HT, Zoghbi HY (2007) Trinucleotide repeat disorders. Annu Rev Neurosci 30:575–621
15. Mirkin SM (2007) Expandable DNA repeats and human disease. Nature 447:932–940
16. Lahue RS, Slater DL (2003) DNA repair and trinucleotide repeat instability. Front Biosci 8:s653–s665
17. Wells RD, Dere R, Hebert ML, Napierala M, Son LS (2005) Advances in mechanisms of genetic instability related to hereditary neurological diseases. Nucleic Acids Res 33:3785–3798
18. Wang G, Christensen LA, Vasquez KM (2006) Z-DNA-forming sequences generate large-scale deletions in mammalian cells. Proc Natl Acad Sci USA 103:2677–2682
19. Adachi M, Tsujimoto Y (1990) Potential Z-DNA elements surround the breakpoints of chromosome translocation within the 5′ flanking region of bcl-2 gene. Oncogene 5:1653–1657
20. Raghavan SC, Lieber MR (2004) Chromosomal translocations and non-B DNA structures in the human genome. Cell Cycle 3:762–768
21. Raghavan SC, Chastain P, Lee JS, Hegde BG, Houston S, Langen R, Hsieh CL, Haworth IS, Lieber MR (2005) Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation. J Biol Chem 280:22749–22760
22. Raghavan SC, Lieber MR (2006) DNA structures at chromosomal translocation sites. Bioessays 28:480–494
23. Wang G, Carbajal S, Vijg J, DiGiovanni J, Vasquez KM (2008) DNA structure-induced genomic instability in vivo. J Natl Cancer Inst 100:1815–1817
24. Kato T, Inagaki H, Yamada K, Kogo H, Ohye T, Kowa H, Nagaoka K, Taniguchi M, Emanuel BS, Kurahashi H (2006) Genetic variation affects de novo translocation frequency. Science 311:971
25. Inagaki H, Ohye T, Kogo H, Kato T, Bolor H, Taniguchi M, Shaikh TH, Emanuel BS, Kurahashi H (2009) Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. Genome Res 19:191–198
26. Emanuel BS (2008) Molecular mechanisms and diagnosis of chromosome 22q11.2 rearrangements. Dev Disabil Res Rev 14:11–18
27. Kurahashi H, Inagaki H, Ohye T, Kogo H, Kato T, Emanuel BS (2006) Palindrome-mediated chromosomal translocations in humans. DNA Repair (Amst) 5:1136–1145
28. Gotter AL, Shaikh TH, Budarf ML, Rhodes CH, Emanuel BS (2004) A palindrome-mediated mechanism distinguishes translocations involving LCR-B of chromosome 22q11.2. Hum Mol Genet 13:103–115
29. Bacolla A, Wells RD (2009) Non-B DNA conformations as determinants of mutagenesis and human disease. Mol Carcinog 48:273–285
30. Smith GR (2008) Meeting DNA palindromes head-to-head. Genes Dev 22:2612–2620
31. Watson J, Hays FA, Ho PS (2004) Definitions and analysis of DNA Holliday junction geometry. Nucleic Acids Res 32:3017–3027
32. Sinden RR, Pettijohn DE (1984) Cruciform transitions in DNA. J Biol Chem 259:6593–6600
33. Oussatcheva EA, Pavlicek J, Sankey OF, Sinden RR, Lyubchenko YL, Potaman VN (2004) Influence of global DNA topology on cruciform formation in supercoiled DNA. J Mol Biol 338:735–743
34. Nag DK, Petes TD (1991) Seven-base-pair inverted repeats in DNA form stable hairpins in vivo in *Saccharomyces cerevisiae*. Genetics 129:669–673
35. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA,

Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

36. Harvey SC (1983) DNA structural dynamics: longitudinal breathing as a possible mechanism for the B in equilibrium Z transition. Nucleic Acids Res 11:4867–4878
37. Peck LJ, Nordheim A, Rich A, Wang JC (1982) Flipping of cloned $d(pCpG)_n \cdot d(pCpG)_n$ DNA sequences from right- to left-handed helical structure by salt, Co(III), or negative supercoiling. Proc Natl Acad Sci USA 79:4560–4564
38. Singleton CK, Klysik J, Stirdivant SM, Wells RD (1982) Left-handed Z-DNA is induced by supercoiling in physiological ionic conditions. Nature 299:312–316
39. Ha SC, Lowenhaupt K, Rich A, Kim YG, Kim KK (2005) Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. Nature 437:1183–1186
40. Htun H, Dahlberg JE (1988) Single strands, triple strands, and kinks in H-DNA. Science 241:1791–1796
41. Wells RD (1988) Unusual DNA structures. J Biol Chem 263:1095–1098
42. Jain A, Wang G, Vasquez KM (2008) DNA triple helices: biological consequences and therapeutic potential. Biochimie 90:1117–1130
43. Majumdar A, Patel DJ (2002) Identifying hydrogen bond alignments in multistranded DNA architectures by NMR. Acc Chem Res 35:1–11
44. Sinden RR, Pytlos-Sinden MJ, Potaman VN (2007) Slipped strand DNA structures. Front Biosci 12:4788–4799
45. Chou SH, Chin KH, Wang AH (2003) Unusual DNA duplex and hairpin motifs. Nucleic Acids Res 31:2461–2474
46. Wang G, Zhao J, Vasquez KM (2009) Methods to determine DNA structural alterations and genetic instability. Methods 48:54–62
47. Pearson CE, Eichler EE, Lorenzetti D, Kramer SF, Zoghbi HY, Nelson DL, Sinden RR (1998) Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. Biochemistry 37:2701–2708
48. Caskey CT, Pizzuti A, Fu YH, Fenwick RG Jr, Nelson DL (1992) Triplet repeat mutations in human disease. Science 256:784–789
49. Benton CS, de Silva R, Rutledge SL, Bohlega S, Ashizawa T, Zoghbi HY (1998) Molecular and clinical studies in SCA-7 define a broad clinical spectrum and the infantile phenotype. Neurology 51:1081–1086
50. Palecek E (1991) Local supercoil-stabilized DNA structures. Crit Rev Biochem Mol Biol 26:151–226
51. Schroth GP, Ho PS (1995) Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. Nucleic Acids Res 23:1977–1983
52. Lafer EM, Moller A, Nordheim A, Stollar BD, Rich A (1981) Antibodies specific for left-handed Z-DNA. Proc Natl Acad Sci USA 78:3546–3550
53. Nordheim A, Pardue ML, Lafer EM, Moller A, Stollar BD, Rich A (1981) Antibodies to left-handed Z-DNA bind to interband regions of *Drosophila* polytene chromosomes. Nature 294:417–422
54. Nordheim A, Lafer EM, Peck LJ, Wang JC, Stollar BD, Rich A (1982) Negatively supercoiled plasmids contain left-handed Z-DNA segments as detected by specific antibody binding. Cell 31:309–318
55. Lafer EM, Sousa R, Ali R, Rich A, Stollar BD (1986) The effect of anti-Z-DNA antibodies on the B-DNA–Z-DNA equilibrium. J Biol Chem 261:6438–6443
56. Agazie YM, Lee JS, Burkholder GD (1994) Characterization of a new monoclonal antibody to triplex DNA and immunofluorescent staining of mammalian chromosomes. J Biol Chem 269:7019–7023
57. Lee JS, Burkholder GD, Latimer LJ, Haug BL, Braun RP (1987) A monoclonal antibody to triplex DNA binds to eucaryotic chromosomes. Nucleic Acids Res 15:1047–1061
58. Agazie YM, Burkholder GD, Lee JS (1996) Triplex DNA in the nucleus: direct binding of triplex-specific antibodies and their effect on transcription, replication and cell growth. Biochem J 316(Pt 2):461–466
59. Ohno M, Fukagawa T, Lee JS, Ikemura T (2002) Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies. Chromosoma 111:201–213
60. Brown JC, Brown BA 2nd, Li Y, Hardin CC (1998) Construction and characterization of a quadruplex DNA selective single-chain autoantibody from a viable motheaten mouse hybridoma with homology to telomeric DNA binding proteins. Biochemistry 37:16338–16348
61. Brown BA 2nd, Li Y, Brown JC, Hardin CC, Roberts JF, Pelsue SC, Shultz LD (1998) Isolation and characterization of a monoclonal anti-quadruplex DNA antibody from autoimmune "viable motheaten" mice. Biochemistry 37:16325–16337
62. Schaffitzel C, Berger I, Postberg J, Hanes J, Lipps HJ, Pluckthun A (2001) In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with *Stylonychia lemnae* macronuclei. Proc Natl Acad Sci USA 98:8572–8577
63. Frappier L, Price GB, Martin RG, Zannis-Hadjopoulos M (1989) Characterization of the binding specificity of two anticruciform DNA monoclonal antibodies. J Biol Chem 264:334–341
64. Sinden RR (1994) Cruciform structures in DNA and triplex DNA in DNA structure and function. Academic, San Diego, pp 160–164 (see also pp 241–242)
65. Raghavan SC, Tsai A, Hsieh CL, Lieber MR (2006) Analysis of non-B DNA structure at chromosomal sites in the mammalian genome. Methods Enzymol 409:301–316
66. Cox R, Mirkin SM (1997) Characteristic enrichment of DNA repeats in different genomes. Proc Natl Acad Sci USA 94:5237–5242
67. Repping S, Skaletsky H, Lange J, Silber S, Van Der Veen F, Oates RD, Page DC, Rozen S (2002) Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. Am J Hum Genet 71:906–922
68. Lobachev KS, Rattray A, Narayanan V (2007) Hairpin- and cruciform-mediated chromosome breakage: causes and consequences in eukaryotic cells. Front Biosci 12:4208–4220
69. Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G (2004) Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly

homologous inverted repeats that contain testes genes. Genome Res 14:1861–1869

70. Khuu P, Sandor M, DeYoung J, Ho PS (2007) Phylogenomic analysis of the emergence of GC-rich transcription elements. Proc Natl Acad Sci USA 104:16528–16533
71. Nordheim A, Rich A (1983) Negatively supercoiled simian virus 40 DNA contains Z-DNA segments within transcriptional enhancer sequences. Nature 303:674–679
72. Oh DB, Kim YG, Rich A (2002) Z-DNA-binding proteins can act as potent effectors of gene expression in vivo. Proc Natl Acad Sci USA 99:16666–16671
73. Wong B, Chen S, Kwon JA, Rich A (2007) Characterization of Z-DNA as a nucleosome-boundary element in yeast *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 104:2229–2234
74. Bacolla A, Collins JR, Gold B, Chuzhanova N, Yi M, Stephens RM, Stefanov S, Olsh A, Jakupciak JP, Dean M, Lempicki RA, Cooper DN, Wells RD (2006) Long homopurine*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. Nucleic Acids Res 34:2663–2675
75. Todd AK, Johnston M, Neidle S (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Res 33:2901–2907
76. Huppert JL, Balasubramanian S (2005) Prevalence of quadruplexes in the human genome. Nucleic Acids Res 33:2908–2916
77. Sundquist WI, Klug A (1989) Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. Nature 342:825–829
78. Williamson JR, Raghuraman MK, Cech TR (1989) Monovalent cation-induced structure of telomeric DNA: the G-quartet model. Cell 59:871–880
79. Panyutin IG, Kovalsky OI, Budowsky EI (1989) Magnesium-dependent supercoiling-induced transition in $(dG)_n \cdot (dC)_n$ stretches and formation of a new G-structure by $(dG)_n$ strand. Nucleic Acids Res 17:8257–8271
80. Huppert JL, Balasubramanian S (2007) G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res 35:406–413
81. Du Z, Zhao Y, Li N (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. Genome Res 18:233–241
82. Du Z, Kong P, Gao Y, Li N (2007) Enrichment of G4 DNA motif in transcriptional regulatory region of chicken genome. Biochem Biophys Res Commun 354:1067–1070
83. Huppert JL, Bugaut A, Kumari S, Balasubramanian S (2008) G-quadruplexes: the beginning and end of UTRs. Nucleic Acids Res 36:6260–6268
84. Sen D, Gilbert W (1990) A sodium–potassium switch in the formation of four-stranded G4-DNA. Nature 344:410–414
85. Moyzis RK, Torney DC, Meyne J, Buckingham JM, Wu JR, Burks C, Sirotkin KM, Goad WB (1989) The distribution of interspersed repetitive DNA sequences in the human genome. Genomics 4:273–289
86. Stallings RL, Torney DC, Hildebrand CE, Longmire JL, Deaven LL, Jett JH, Doggett NA, Moyzis RK (1990) Physical mapping of human chromosomes by repetitive sequence fingerprinting. Proc Natl Acad Sci USA 87:6218–6222
87. Krontiris TG (1995) Minisatellites and human disease. Science 269:1682–1683
88. Bacolla A, Wojciechowska M, Kosmider B, Larson JE, Wells RD (2006) The involvement of non-B DNA structures in gross chromosomal rearrangements. DNA Repair (Amst) 5:1161–1170
89. Bacolla A, Larson JE, Collins JR, Li J, Milosavljevic A, Stenson PD, Cooper DN, Wells RD (2008) Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. Genome Res 18:1545–1553
90. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945
91. Eichler EE, Clark RA, She X (2004) An assessment of the sequence gaps: unfinished business in a finished human genome. Nat Rev Genet 5:345–354
92. Repping S, van Daalen SK, Brown LG, Korver CM, Lange J, Marszalek JD, Pyntikova T, van der Veen F, Skaletsky H, Page DC, Rozen S (2006) High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat Genet 38:463–467
93. Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37:1–13
94. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320:539–543
95. Venkatasubramanian G (2009) Triplex DNA, human evolution and schizophrenia. Acta Neuropsychiatr 21:100–101
96. Alba MM, Guigo R (2004) Comparative analysis of amino acid repeats in rodents and humans. Genome Res 14:549–554
97. Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Banda MG, Whisstock JC (2005) Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res 15:537–551
98. Schaafsma D, Roscioni SS, Meurs H, Schmidt M (2008) Monomeric G-proteins as signal transducers in airway physiology and pathophysiology. Cell Signal 20:1705–1714
99. Burridge K, Wennerberg K (2004) Rho and Rac take center stage. Cell 116:167–179
100. Zhao Y, Du Z, Li N (2007) Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. FEBS Lett 581:1951–1956
101. Rawal P, Kummarasetti VB, Ravindran J, Kumar N, Halder K, Sharma R, Mukerji M, Das SK, Chowdhury S (2006) Genome-wide prediction of G4 DNA as regulatory motifs: role in *Escherichia coli* global regulation. Genome Res 16:644–655
102. Qin Y, Hurley LH (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. Biochimie 90:1149–1171
103. Siddiqui-Jain A, Grand CL, Bearss DJ, Hurley LH (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proc Natl Acad Sci USA 99:11593–11598
104. Hurley LH, Von Hoff DD, Siddiqui-Jain A, Yang D (2006) Drug targeting of the c-MYC promoter to repress gene expression via a G-quadruplex silencer element. Semin Oncol 33:498–512
105. Aitken RJ, Marshall Graves JA (2002) The future of sex. Nature 415:963
106. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature 423:873–876
107. Bhowmick BK, Satta Y, Takahata N (2007) The origin and evolution of human ampliconic gene families and ampliconic structure. Genome Res 17:441–450
108. Hughes JF, Skaletsky H, Pyntikova T, Minx PJ, Graves T, Rozen S, Wilson RK, Page DC (2005) Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. Nature 437:100–103

109. Kolb J, Chuzhanova NA, Hogel J, Vasquez KM, Cooper DN, Bacolla A, Kehrer-Sawatzki H (2009) Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. Chromosome Res 7:469–483

110. Losch FO, Bredenbeck A, Hollstein VM, Walden P, Wrede P (2007) Evidence for a large double-cruciform DNA structure on the X chromosome of human and chimpanzee. Hum Genet 122:337–343

111. Kurahashi H, Inagaki H, Kato T, Hosoba E, Kogo H, Ohye T, Tsutsumi M, Bolor H, Tong M, Emanuel BS (2009) Impaired DNA replication prompts deletions within palindromic sequences, but does not induce translocations in human cells. Hum Mol Genet 18:3397–3406

112. Sinden RR, Bat O, Kramer PR (1999) Psoralen cross-linking as probe of torsional tension and topological domain size in vivo. Methods 17:112–124

113. Pearson CE, Nichol Edamura K, Cleary JD (2005) Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 6:729–742

114. Mirkin EV, Mirkin SM (2007) Replication fork stalling at natural impediments. Microbiol Mol Biol Rev 71:13–35

115. Wang G, Vasquez KM (2009) Models for chromosomal replication-independent non-B DNA structure-induced genetic instability. Mol Carcinog 48:286–298

116. Freudenreich CH (2007) Chromosome fragility: molecular mechanisms and cellular consequences. Front Biosci 12:4911–4924

117. Collins NS, Bhattacharyya S, Lahue RS (2007) Rev1 enhances CAG·CTG repeat stability in *Saccharomyces cerevisiae*. DNA Repair (Amst) 6:38–44

118. Voineagu I, Narayanan V, Lobachev KS, Mirkin SM (2008) Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. Proc Natl Acad Sci USA 105:9936–9941

119. Zhang H, Freudenreich CH (2007) An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in *S. cerevisiae*. Mol Cell 27:367–379

120. Balakumaran BS, Freudenreich CH, Zakian VA (2000) CGG/CCG repeats exhibit orientation-dependent instability and orientation-independent fragility in *Saccharomyces cerevisiae*. Hum Mol Genet 9:93–100

121. Panigrahi GB, Cleary JD, Pearson CE (2002) In vitro (CTG)*(CAG) expansions and deletions by human cell extracts. J Biol Chem 277:13926–13934

122. Kang S, Jaworski A, Ohshima K, Wells RD (1995) Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in *E. coli*. Nat Genet 10:213–218

123. Freudenreich CH, Stavenhagen JB, Zakian VA (1997) Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. Mol Cell Biol 17:2090–2098

124. Kim HM, Narayanan V, Mieczkowski PA, Petes TD, Krasilnikova MM, Mirkin SM, Lobachev KS (2008) Chromosome fragility at GAA tracts in yeast depends on repeat orientation and requires mismatch repair. EMBO J 27:2896–2906

125. Hebert ML, Spitz LA, Wells RD (2004) DNA double-strand breaks induce deletion of CTG·CAG repeats in an orientation-dependent manner in *Escherichia coli*. J Mol Biol 336:655–672

126. Mitas M (1997) Trinucleotide repeats associated with human disease. Nucleic Acids Res 25:2245–2254

127. Pearson CE, Tam M, Wang YH, Montgomery SE, Dar AC, Cleary JD, Nichol K (2002) Slipped-strand DNAs formed by long (CAG)*(CTG) repeats: slipped-out repeats and slip-out junctions. Nucleic Acids Res 30:4534–4547

128. Kovtun IV, McMurray CT (2008) Features of trinucleotide repeat instability in vivo. Cell Res 18:198–213

129. Owen BA, Yang Z, Lai M, Gajec M, Badger JD 2nd, Hayes JJ, Edelmann W, Kucherlapati R, Wilson TM, McMurray CT (2005) $(CAG)_n$-hairpin DNA binds to Msh2–Msh3 and changes properties of mismatch recognition. Nat Struct Mol Biol 12:663–670

130. Manley K, Shirley TL, Flaherty L, Messer A (1999) Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nat Genet 23:471–473

131. Lin Y, Dion V, Wilson JH (2006) Transcription promotes contraction of CAG repeat tracts in human cells. Nat Struct Mol Biol 13:179–180

132. Voineagu I, Surka CF, Shishkin AA, Krasilnikova MM, Mirkin SM (2009) Replisome stalling and stabilization at CGG repeats, which are responsible for chromosomal fragility. Nat Struct Mol Biol 16:226–228

133. Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, Resnick MA (2000) Inverted Alu repeats unstable in yeast are excluded from the human genome. EMBO J 19:3822–3830

134. Spiro C, Pelletier R, Rolfsmeier ML, Dixon MJ, Lahue RS, Gupta G, Park MS, Chen X, Mariappan SV, McMurray CT (1999) Inhibition of FEN-1 processing by DNA secondary structure at trinucleotide repeats. Mol Cell 4:1079–1085

135. Refsland EW, Livingston DM (2005) Interactions among DNA ligase I, the flap endonuclease and proliferating cell nuclear antigen in the expansion and contraction of CAG repeat tracts in yeast. Genetics 171:923–934

136. Bhattacharyya S, Lahue RS (2005) Srs2 helicase of *Saccharomyces cerevisiae* selectively unwinds triplet repeat DNA. J Biol Chem 280:33311–33317

137. Kerrest A, Anand RP, Sundararajan R, Bermejo R, Liberi G, Dujon B, Freudenreich CH, Richard GF (2009) SRS2 and SGS1 prevent chromosomal breaks and stabilize triplet repeats by restraining recombination. Nat Struct Mol Biol 16:159–167

138. Daee DL, Mertz T, Lahue RS (2007) Postreplication repair inhibits CAG·CTG repeat expansions in *Saccharomyces cerevisiae*. Mol Cell Biol 27:102–110

139. Moe SE, Sorbo JG, Holen T (2008) Huntingtin triplet-repeat locus is stable under long-term Fen1 knockdown in human cells. J Neurosci Methods 171:233–238

140. van den Broek WJ, Nelen MR, van der Heijden GW, Wansink DG, Wieringa B (2006) Fen1 does not control somatic hypermutability of the $(CTG)_n$*$(CAG)_n$ repeat in a knock-in mouse model for DM1. FEBS Lett 580:5208–5214

141. Chong SS, McCall AE, Cota J, Subramony SH, Orr HT, Hughes MR, Zoghbi HY (1995) Gametic and somatic tissue-specific heterogeneity of the expanded SCA1 CAG repeat in spinocerebellar ataxia type 1. Nat Genet 10:344–350

142. Telenius H, Kremer B, Goldberg YP, Theilmann J, Andrew SE, Zeisler J, Adam S, Greenberg C, Ives EJ, Clarke LA et al (1994) Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. Nat Genet 6:409–414

143. Kovtun IV, McMurray CT (2001) Trinucleotide expansion in haploid germ cells by gap repair. Nat Genet 27:407–411

144. Trujillo KM, Sung P (2001) DNA structure-specific nuclease activities in the *Saccharomyces cerevisiae* Rad50*Mre11 complex. J Biol Chem 276:35458–35464

145. Nag DK, Fasullo M, Dong Z, Tronnes A (2005) Inverted repeat-stimulated sister-chromatid exchange events are RAD1-independent but reduced in a msh2 mutant. Nucleic Acids Res 33:5243–5249

146. Kirkpatrick DT, Petes TD (1997) Repair of DNA loops involves DNA-mismatch and nucleotide-excision repair proteins. Nature 387:929–931

147. Parniewski P, Bacolla A, Jaworski A, Wells RD (1999) Nucleotide excision repair affects the stability of long transcribed (CTG*CAG) tracts in an orientation-dependent manner in *Escherichia coli*. Nucleic Acids Res 27:616–623

148. Pelletier R, Farrell BT, Miret JJ, Lahue RS (2005) Mechanistic features of CAG*CTG repeat contractions in cultured cells revealed by a novel genetic assay. Nucleic Acids Res 33:5667–5676

149. Panigrahi GB, Lau R, Montgomery SE, Leonard MR, Pearson CE (2005) Slipped (CTG)*(CAG) repeats can be correctly repaired, escape repair or undergo error-prone repair. Nat Struct Mol Biol 12:654–662

150. Savouret C, Garcia-Cordier C, Megret J, te Riele H, Junien C, Gourdon G (2004) MSH2-dependent germinal CTG repeat expansions are produced continuously in spermatogonia from DM1 transgenic mice. Mol Cell Biol 24:629–637

151. Wang G, Vasquez KM (2007) Z-DNA, an active element in the genome. Front Biosci 12:4424–4438

152. 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

153. Rolfsmeier ML, Dixon MJ, Lahue RS (2000) Mismatch repair blocks expansions of interrupted trinucleotide repeats in yeast. Mol Cell 6:1501–1507

154. Savouret C, Brisson E, Essers J, Kanaar R, Pastink A, te Riele H, Junien C, Gourdon G (2003) CTG repeat instability and size variation timing in DNA repair-deficient mice. EMBO J 22:2264–2273

155. Hebert ML, Wells RD (2005) Roles of double-strand breaks, nicks and gaps in stimulating deletions of CTG·CAG repeats by intramolecular DNA repair. J Mol Biol 353:961–979

156. Pollard LM, Bourn RL, Bidichandani SI (2008) Repair of DNA double-strand breaks within the (GAA*TTC)$_n$ sequence results in frequent deletion of the triplet-repeat sequence. Nucleic Acids Res 36:489–500

157. Marcadier JL, Pearson CE (2003) Fidelity of primate cell repair of a double-strand break within a (CTG)·(CAG) tract. Effect of slipped DNA structures. J Biol Chem 278:33848–33856

158. Downing B, Morgan R, VanHulle K, Deem A, Malkova A (2008) Large inverted repeats in the vicinity of a single double-strand break strongly affect repair in yeast diploids lacking Rad51. Mutat Res 645:9–18

159. VanHulle K, Lemoine FJ, Narayanan V, Downing B, Hull K, McCullough C, Bellinger M, Lobachev K, Petes TD, Malkova A (2007) Inverted DNA repeats channel repair of distant double-strand breaks into chromatid fusions and chromosomal rearrangements. Mol Cell Biol 27:2601–2614

160. Jakupciak JP, Wells RD (2000) Gene conversion (recombination) mediates expansions of CTG·CAG repeats. J Biol Chem 275:40003–40013

161. Tartier L, Michalik V, Spotheim-Maurizot M, Rahmouni AR, Sabattier R, Charlier M (1994) Radiolytic signature of Z-DNA. Nucleic Acids Res 22:5565–5570

162. Ribeiro DT, Madzak C, Sarasin A, Di Mascio P, Sies H, Menck CF (1992) Singlet oxygen induced DNA damage and mutagenicity in a single-stranded SV40-based shuttle vector. Photochem Photobiol 55:39–45

163. Lagravere C, Malfoy B, Leng M, Laval J (1984) Ring-opened alkylated guanine is not repaired in Z-DNA. Nature 310:798–800

164. Boiteux S, Costa de Oliveira R, Laval J (1985) The *Escherichia coli* O6-methylguanine-DNA methyltransferase does not repair promutagenic O6-methylguanine residues when present in Z-DNA. J Biol Chem 260:8711–8715