

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:<https://orca.cardiff.ac.uk/id/eprint/110308/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Bacolla, Albino, Cooper, David Neil , Vasquez, Karen M. and Tainer, John A. 2018. Non-B DNA structure and mutations causing human genetic disease. eLS, John Wiley & Sons, (10.1002/9780470015902.a0022657.pub2)

Publishers page: <http://dx.doi.org/10.1002/9780470015902.a0022657.p...>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



1 eLS
2 Non-B DNA Structure and Mutations Causing Human Disease

3 Albino Bacolla¹,
4 David N Cooper²,
5 Karen M Vasquez³
6 John A Tainer¹

7 ¹ The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

8 ² Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK

9 ³ The University of Texas at Austin, Austin, Texas, USA

10 Published Online: 18 OCT 2010

11

12 Advanced Article

13 (Advanced articles are aimed at advanced undergraduates, graduate students, postgraduates, and
14 researchers reading outside their field of expertise.)

15

16 Abstract

17 In addition to the canonical right-handed double helix, several noncanonical deoxyribonucleic acid
18 (DNA) secondary structures have been characterised, including quadruplexes, triplexes,
19 slipped/hairpins, Z-DNA and cruciforms collectively termed non-B DNA. The formation of non-B DNA
20 is mediated by repetitive sequence motifs, such as G-rich sequences, purine/pyrimidine tracts, direct
21 (tandem) repeats, alternating purine–pyrimidines and inverted repeats, respectively. Such repeats
22 are abundant in the human genome and non-B DNA has been found at specific genomic locations,
23 supporting a role in gene regulation, RNA translation and protein function. Repetitive motifs are also
24 found at sites of chromosomal alterations associated with both human genetic disease and cancer.
25 Characterised by an inherent capacity to expand spontaneously, such sequences are not only known
26 to cause >30 neurological diseases but may also contribute to human disease susceptibility. The
27 formation of non-B DNA structures is believed to promote genomic alterations by impeding efficient
28 DNA replication, transcription, and repair.

29 Key Concepts:

30 The structure of DNA is polymorphic as well as its sequence; in addition to the canonical right-
31 handed double helix (B-DNA), repetitive sequences can also adopt alternative (non-B DNA)
32 conformations such as quadruplexes, triplexes, slipped/hairpins, Z-DNA and cruciforms.

33 Repetitive DNA sequences are found at locations within many human genes that suggest they can
34 either affect transcription or alternatively encode homopolymeric amino acid runs that could be
35 important for either protein–protein or protein–DNA/RNA interactions.

36 G4 and Z-DNA structures have been detected in cells through specific antibodies, mostly in
37 correspondence of actively transcribed genes and, in the case of G4, at telomeres.

1 Copy number variation (CNV) is a form of genetic alteration that, by involving thousands of loci in
2 the genome, contributes to human individuality.

3 Repetitive sequences capable of forming non-B DNA are found at sites of chromosomal breaks, CNVs
4 and other rearrangements such as translocations, deletions and gene conversion events, which can
5 contribute to human genetic disease and cancer.

6 The recurrent translocation t(22;11) events associated with Emanuel syndrome are mediated by
7 cruciform structures that occur at inverted repeats.

8 Tandem repeats (microsatellites) may expand within gene sequences, contributing to more than 30
9 neurological diseases; present in variable number in genes in the population, they may contribute to
10 human disease susceptibility.

11 An increasing number of enzymes are being discovered that resolve non-B DNA structures, and
12 whose mutations lead to genomic instability and human disease.

13 lncRNAs repress gene expression by forming triplex structures with their target duplex DNA.

14 Non-B DNA structures stimulate mutations via mechanisms that alter DNA synthesis, transcription and
15 repair.

16 Keywords:

17 non-B DNA;

18 microsatellites;

19 copy number variation (CNV);

20 triplet repeat diseases;

21 polyglutamine expansion;

22 translocations;

23 DNA repair;

24 DNA replication;

25 double strand breaks (DSB);

26 gene expression regulation

27 cancer genomes

28 helicases

29 RAN translation

30 lncRNA

31

32 Abstract: Please expand the abstract to 120 – 150 words

33 Key Concepts: Please include up to ten key concepts in your manuscript. Key concepts should sum
34 up the essential ideas in the manuscript, rather than listing the article contents. Key concepts should

1 not be confused with key words (which are for indexing purposes) or glossary. A key concept should
2 be described in a short sentence and should be presented in bullet style, e.g.

3 Animal behaviorists must participate in conservation planning to protect the future of biodiversity.

4 Lipid bilayers provide the fundamental architecture of biological membranes.

5

6 Introduction

7 Soon after the discovery that deoxyribonucleic acid (DNA) was an antiparallel right-handed double
8 helix, work on synthetic single-stranded DNA molecules of defined sequence composition revealed
9 the formation of a three-stranded structure, in addition to the expected duplex. This implied the
10 existence of more than one type of DNA conformation. During the subsequent 60 years, the
11 repertoire of conformations in which synthetic DNA molecules of repeating sequence composition
12 were found to assemble, increased steadily. To date, several unique DNA structures, quite distinct
13 from the canonical B-form, have been characterised, including left-handed Z-DNA, cruciforms,
14 looped-out or slipped folds, parallel DNA, triplexes, quadruplexes and higher-order arrangements.
15 These conformations are collectively called non-B DNA.

16 In parallel to these biophysical investigations, DNA sequencing revealed the existence in
17 chromosomes of several different types of repetitive motifs known to adopt non-B forms in vitro,
18 spurring speculation as to their biological function. Pioneering work in bacteria suggested a role for
19 non-B DNA-forming sequences in mediating chromosomal deletions in vivo. However, it was not
20 until comparatively recently that the concept of DNA secondary structure as a promoter of genomic
21 rearrangements acquired broad support. Critical to these developments was the discovery of
22 microsatellite repeat diseases (MRDs), a novel class of neurological disorders caused by the
23 expansion of triplet (and other) microsatellite repeats (Brouwer et al., 2009; Lee and Cooper, 2009;
24 Lopez Castel et al., 2010; Wells and Ashizawa, 2006).

25 The completion of the human and other mammalian genome-sequencing projects made it possible
26 to explore the frequencies and locations of chromosomal sites containing non-B DNA-forming
27 sequences, as well as their evolutionary conservation in orthologous genomes. From a number of
28 studies (reviewed in Zhao et al., 2010) it could be concluded that such sites occur much more
29 frequently than expected by chance alone and that different non-B DNA-forming motifs are found in
30 association with different genomic regions. Hence, a novel view of genome structure and function
31 emerged in which repetitive DNA, once regarded as 'junk DNA', participates in the regulation of
32 genes, telomere maintenance, RNA metabolism, protein function and genome stability.

33 The human genome sequencing project has also led to the realisation that many individual genomes
34 exist (in contrast to the early concept of a nearly identical standard genome common to all human
35 beings) in which a large number of gross variations involving duplicated and deleted regions (copy
36 number variations or CNVs) comprising large (>1 kb in length, on average) segments of DNA
37 distinguish one individual from another. More recently, chromatin immunoprecipitation techniques
38 associated with high throughput DNA sequencing has made it possible to map directly non-B DNA
39 structures in cells, and to begin identifying the protein complexes with which these structures
40 interact and elicit their biological functions. Overall, non-B DNA structures are emerging as a
41 powerful tool used by cells to regulate gene expression, RNA translation, and most likely other
42 processes. At the same time, an increasing large repertoire of enzymes that recognize and resolve
43 these structures are being identified, whose general role is to prevent their deleterious effects on

1 genome stability. Hence, non-B DNA (and RNA) structures may be regarded as a double edged sword,
2 critical for biological function and at the same time a threat to genome integrity.

3 Herein, we review the main forms of non-B DNA structure, outline their distributions in the human
4 genome, and present selected work on the relationship between repetitive DNA sequences and
5 human disease.

6 Non-B DNA Structures

7 To date, at least five types of non-B DNA structures have been associated with human disease:
8 cruciform, triplex, slipped/hairpin DNA, quadruplex and Z-DNA (Figure 1). Most DNA sequences in
9 the human genome exist in the B-form. However, repetitive sequences may also adopt alternative
10 conformations as a consequence of multiple hydrogen bonding interactions between bases or, as in
11 the case of Z-DNA, rotational freedom about the N-glycosidic bond of G residues. Triplex and
12 quadruplex DNA also rely on additional hydrogen bonding donor and acceptor groups existing in
13 purines (A and G).

14 <FIGURE 1 NEAR HERE>

15 Cruciform DNA

16 Inverted repeats (IR), defined as a series of nucleotides followed on the same strand by their
17 complementary bases and usually separated by a spacer, are required for cruciform formation
18 (Figure 1a). The term 'complementary' here refers to the fact that A always pairs with T whereas G
19 always pairs with C in B-form DNA (so-called 'Watson–Crick' base-pairing).

20 The IR symmetry makes it possible for the bases along the same strand of DNA to pair with each
21 other, rather than with the complementary strand, thereby giving rise to a cross-shaped structure
22 (Figure 1a). In solution, two interconvertible conformations have been observed: an extended
23 conformation, in which each cruciform arm occupies the vertex of a simple tetrahedron, and a
24 closed conformation, in which the two pairs of arms lie almost parallel to each other. The closed
25 conformation is favoured at physiological salt concentrations because of the shielding of negatively
26 charged phosphates. Thus, it is considered a good approximation of cruciform (and Holliday
27 junction) structures in vivo.

28 Triplex DNA (H-DNA)

29 Triplex DNA is a three-stranded structure in which a third (DNA or ribonucleic acid (RNA)) strand
30 binds to B-DNA by occupying its wide major groove. Watson–Crick hydrogen bonds are not altered in
31 H-DNA; rather, purine bases along one strand of duplex DNA engage the incoming third strand
32 through Hoogsteen-hydrogen bonds using available exocyclic groups not involved in Watson–Crick
33 interactions. This third strand may originate from a nearby sequence on the same molecule
34 (intramolecular triplex) or from separate molecules (intermolecular triplex), comprising either RNA,
35 DNA or exogenously added triplex-forming oligonucleotides (TFOs). Hence, triplex DNA requires a
36 succession of purines on the same strand of DNA. Four types of Hoogsteen-hydrogen bonds typically
37 occur: A:T, A:A, G:C⁺ (C⁺, protonated base) and G:G (Figure 1b and c), which yield two types of
38 triplexes: a YRY type, in which the third strand is pyrimidine-rich (Figure 1c), and an RRY type, in
39 which the third strand is purine-rich (Figure 1b). In intramolecular triplexes, a mirror repeat (MR)
40 symmetry is required in duplex DNA, whereby bases from one repeat separate from the
41 complementary strand and fold back onto the second repeat to engage in Hoogsteen pairing. This
42 reaction requires energy, which can be provided by negative supercoiling. If it is the pyrimidine-rich

1 strand that engages in Hoogsteen base-pairing (such as occurs at low pH), the third strand runs
2 parallel to the purine-rich accepting strand; however, if it is the purine-rich strand that engages in
3 Hoogsteen pairing (at neutral pH), the third strand runs antiparallel to the purine-rich accepting
4 partner, as shown in Figure 1a. Under in vitro conditions that mimic the molecular crowding of the
5 cell environment, DDR (D = DNA; R = RNA) triplexes appear to be most stable.

6 Slipped/hairpin DNA

7 The reiteration of bases, such as CTGCTGCTG, or direct repeats (DR), provides the basis for the
8 bulging out of small loops in duplex DNA when, after separation of the complementary strands, the
9 repeat array reanneals out-of-register. Thus, slipped DNA occurs during DNA replication and
10 transcription, two processes that necessarily entail strand separation. An alternative source of small
11 loops in DNA (particularly at mononucleotide runs, such as AAA) originates from stuttering or
12 skipping during DNA replication. Looped-out sequences may fold into double helices (hairpins)
13 stabilised by mismatches, such as T•T base pairs, in addition to the Watson–Crick A•T and G•C base
14 pairs along the hairpins.

15 Quadruplex (tetraplex, G4) DNA

16 G4 DNA has received considerable attention during the past few years, in part because human
17 telomeres, whose function is often dysregulated in cancer, are composed exclusively of hexameric
18 (TTAGGG)_n repeat sequences capable of forming this type of non-B DNA structure. For quadruplexes
19 to form, the sequence pattern required comprises four closely spaced sets of 2–4 Gs, such as the 3-
20 set GGG(n₁)GGG(n₂)GGG(n₃)GGG, where n₁, n₂ and n₃ represent 1–7 bases of any kind (loop). The
21 basic unit in G4 DNA, a G-tetrad or G-quartet, comprises a planar array of four guanines (one from
22 each of the four sets) connected to each other through Hoogsteen-type hydrogen bonds (Figure 1d).
23 A core of at least two (three in the example shown in Figure 1a) G-quartets stack on top of one
24 another, stabilised by cation coordination (potassium ion is most effective) between any two stacks
25 while the loops provide strand connectivity along the edges of the stacked G-quartets. A high degree
26 of structural polymorphism has been revealed, in which loops connect in a lateral, diagonal or chain
27 reversal fashion, strands run parallel or antiparallel to one another, and guanine residues may adopt
28 either the syn or the anti N-glycosidic conformation (Neidle, 2009; Figure 1a). Thus, competing
29 conformations usually form in solution. The most common isomer, as evidenced by nuclear magnetic
30 resonance (NMR) and X-ray crystallography, is represented by the (3+1) mixed topology,
31 characterised by one chain reversal and two lateral loops and by a three syn plus one anti-
32 arrangement of guanines per G-tetrad (Neidle, 2009).

33 Z-DNA

34 Z-DNA is unusual among the non-B DNA structures in that strandedness is reversed by the rotation
35 from the anti (in right-handed B-DNA) to the syn (in left-handed Z-DNA) conformation of every G
36 residue within tracts of alternating GY (Y, pyrimidine) sequences, such as (GC)_n or (GC)_m(GT)_n. In
37 sharp contradistinction to B-DNA, which possesses both a major and a minor groove, Z-DNA is
38 characterised by a single deep and narrow groove and an overall tube-like shape. X-ray
39 crystallography has also shown that the base pairs located at the junctions between the B- and Z-
40 sections (B-Z junctions, Figure 1a) have the tendency to flip out of the double-helix, thereby
41 providing a substrate for base modification or cleavage. See also DNA Structure; DNA Structure:
42 Sequence Effects; DNA Structure: Sequence Effects; Supercoiled DNA: Structure; Macromolecular
43 Interactions: Aptamers; Base Pairing in DNA: Unusual Patterns.

1 Probing DNA Structures Genome-wide

2 Substantial progress has been made in recent years in mapping non-B DNA structures throughout
3 the human genome using chromatin immunoprecipitation followed by deep sequencing, i.e. ChIP-
4 seq. G4 DNA was probed in a human epidermal keratinocyte cell line by a G4 structure-specific
5 antibody (Hansel-Hertsch et al., 2016), which revealed >10,000 high-confidence peaks (i.e. G4
6 structures), 98% of which coincided with nucleosome-depleted and accessible chromatin regions.
7 G4-rich chromatin displayed higher transcriptional activity than similar regions devoid of G4 DNA,
8 implying that non-B DNA served to enhance gene transcription on a genome-wide scale. A parallel
9 G4 DNA-specific antibody was also employed in ChIP-seq experiments to demonstrate the
10 enrichment of G4 structures at human telomeres (Liu et al. 2016).

11 Z-DNA has been probed using a synthetic peptide, termed Zaa, consisting of two Z α domains from
12 the human double-stranded RNA-specific adenosine deaminase (ADAR), which binds Z-DNA with
13 high selectivity, further attached to FLAG, a protein tag recognized by FLAG-specific antibodies. A
14 total of 391 high-confidence Zaa binding peaks were identified in HeLa cells, mostly near the
15 transcription start sites of actively transcribed genes (Shin et al., 2016). A recent and biologically
16 important extension in the field has been the identification of triplexes formed between duplex DNA
17 and single-stranded RNA from long noncoding RNAs (lncRNAs) using ChIP assays and a triplex-
18 specific antibody (Mondal et al., 2015).

19 Besides direct probing of non-B DNA in cells, earlier bioinformatic analyses shed light on the location
20 of non-B DNA-forming repeats in the human and other genomes, which revealed unexpected
21 complexities. Large IRs (>100 kb) are present on sex chromosomes, with male-specific genes and
22 gene families essential for male fertility located at symmetrical positions along the IR arms (Table 1;
23 Skaletsky et al., 2003). The maintenance of gene function, particularly for the Y-chromosome that
24 lacks a homologue for recombination, is believed to depend on the formation of large cruciform
25 structures by IR sequences, which potentiate the correction of mutations and double-strand breaks
26 by intrachromosomal recombination (Lange et al., 2009; Zhao et al., 2010).

27 <TABLE 1 NEAR HERE>

28

29 Long (≥ 250 bases in length) runs of homopurines•homopyrimidines were found in introns of 228
30 genes (reviewed in Zhao et al., 2010), mostly encoding proteins with a function in cell
31 communication and synaptic transmission of the nerve impulse (Table 1). These gene classes are also
32 enriched in GGAA, GAAA and GGGG tetranucleotide repeats, which have the capacity to form stable
33 triplexes. Although these types of genes are generally weakly transcribed, a high proportion of them
34 is preferentially expressed in the brain.

35 The distribution of short tandem repeats (slipped/hairpin DNA) in protein-coding sequences is
36 dominated by triplet repeats encoding homopolymeric runs of specific amino acids, such as
37 polyglutamine, in transcription factors and gene-regulatory proteins that bind DNA and RNA (Table
38 1). Homopolymeric runs of amino acids are known to play critical roles in protein–protein and
39 protein–DNA/RNA interactions and their number at specific loci appears to increase as one
40 progresses from the genomes of simpler species to the more complex. Thus, DNA slippage and
41 hairpin/loop formation may have been exploited over evolutionary time as a means to acquire, or
42 fine-tune, protein function. See also Genetic Variation: Polymorphisms and Mutations; Next
43 Generation Sequencing Technologies and Their Applications; Advances in Next Generation

1 Sequencing Technologies and Cancer Epigenomics; Long Noncoding RNAs and Cancer; Long
2 Noncoding RNAs and Tumorigenesis; Chromosome Y; Y Chromosome; Disordered Proteins; Protein
3 Aggregation and Human Disorders; Protein Disorder and Human Genetic Disease.

4

5 DNA/RNA Structure, Phenotypic Variation and Human Disease

6 A number of studies have implicated the formation of non-B DNA conformations as a source of
7 genomic rearrangements causing human genetic disease, including Fabry disease (Kornreich et al.,
8 1990), mental retardation (Bonaglia et al., 2009; Rooms et al., 2007), ornithine transcarbamylase
9 deficiency (Quental et al., 2009), blepharophimosis-ptosis-epicanthus inversus syndrome (Verdin et
10 al., 2013), uniparental disomy 14(mat) (Bena et al., 2010) and spermatogenic failure among others
11 (reviewed in Bacolla and Wells, 2004). As an example, Emanuel syndrome (MIM #609029),
12 characterised by severe mental retardation, facial abnormalities and heart and kidney defects, is
13 caused by the inheritance of a supernumerary der(22) chromosome from a parent carrying a
14 constitutional translocation between chromosomes 11 and 22 (t(11;22)(q23;q11)) (Figure 2). Cloning
15 of the genomic regions involved in the translocation revealed that the breakpoints typically occurred
16 within narrow loci on both chr11 and chr22, at the centre of large (~450 bp on chr11 and ~590 bp on
17 chr22) IR structures comprising almost exclusively A and T bases. These recurring breaks, at the
18 centre of specific IRs termed PATRR11 and PATRR22 (palindromic AT-rich regions) respectively, are
19 consistent with the formation of large cruciform structures on both chromosomes (Figure 2). The
20 conclusion is supported by the following observations (Kurahashi et al., 2010). First, the PATRR22
21 sequence was shown to be both polymorphic and intrinsically unstable in the general population,
22 such that deletions and duplications reducing or disrupting IR symmetry were commonly observed.
23 Analyses of t(11;22) frequencies in sperm cells from healthy individuals yielded an estimate of
24 $\sim 1.5 \times 10^{-5}$ for the full-length IR chromosomes, but an ~ 10 -fold reduction for those chromosomes in
25 which IR symmetry was disrupted, implying that cruciform structures were responsible for
26 promoting the translocation event. Second, fluorescence in situ hybridisation indicated that the
27 22q11 cluster was involved in additional translocations, including 17q11, 4q35.1, 1p21.2 and 8q24.1.
28 In all cases, repetitive sequences with IR symmetry were detected on the partner chromosome, with
29 translocation frequencies decreasing with decreasing length of the IR sequences. Thus, size (and
30 hence stability) of cruciform structures correlates with the likelihood of chromosomal breaks. Taken
31 together, these data support the conclusion that large cruciforms are extruded from the PATRR11
32 and PATRR22 sequences, which are then cleaved at the centre, generating double-stranded broken
33 ends that are sealed, yielding the der(22) and der(11) chromosomes (Figure 2).

34 <FIGURE 2 NEAR HERE>

35 The number of identified human disorders associated with gains (duplications) and losses (deletions)
36 of genetic material, types of mutation once believed to occur only rarely (Perry et al., 2008), has
37 increased considerably over the past years (Zhang et al., 2010). The application of aCGH (Conrad et
38 al., 2010; Perry et al., 2008) to patients afflicted with either single-gene disorders, such as the CFTR
39 gene associated with cystic fibrosis (Quemener et al., 2010) and the NRXN1 gene linked to autism
40 spectrum disorders (Chen et al., 2013), or multigene disorders (Vissers et al., 2009; Zhang et al.,
41 2010), indicated that CNVs can also be involved in these conditions. Genome-wide studies support
42 the conclusion that CNVs affect thousands of loci, contributing in all likelihood to the myriad
43 phenotypic differences observed between individuals (Conrad et al., 2010; Perry et al., 2008).
44 Bioinformatic analyses indicate that the density of repetitive DNA sequence motifs capable of

1 adopting non-B DNA conformations is significantly higher at CNV breakpoints than the genome
2 average, implying that the formation of local DNA secondary structure may represent a common
3 mechanism for CNV-mediated deletions, duplications and inversions (Conrad et al., 2010; Perry et
4 al., 2008; Vissers et al., 2009; Chen et al., 2013)...

5 A comprehensive meta-analysis (Chuzhanova et al., 2009) of the DNA sequences flanking the regions
6 of DNA exchange in 27 known disease-associated gene conversion events indicated that non-B DNA-
7 forming sequences, particularly IRs, occurred more frequently than expected by chance alone.
8 Hence, at least some gene conversion events may be initiated by DSBs at sites of non-B DNA
9 structure. Non-B conformations have also been found to occur more often than expected by chance
10 at sites of microlesions (base-pair substitutions and insertions and deletions <21 bp in length) from a
11 large meta-analysis of >83,000 pathological mutations associated with human inherited disease
12 (Kamat et al., 2015). G4-forming sequences were also found to be overrepresented at deletion
13 breakpoints in mitochondrial DNA (Dong et al., 2014). Lymphocytes isolated from patients with mild
14 cognitive impairment showed higher numbers of G4 DNA foci than healthy controls, and indeed G4
15 DNA count by immunofluorescence has been proposed as a biomarker for cognitive disorders, such
16 as Alzheimer's disease (Francois M et al., 2016). Cockayne syndrome (CS), a fatal neurodegenerative
17 disease characterized by accelerated aging, impaired growth, hypersensitivity to sunlight, is caused
18 by mutations in either the ERCC8 (CSA) or ERCC6 (CSB) genes; CS has recently been associated with
19 the accumulation of non-B structures in mitochondrial DNA (Scheibye-Knudsen et al., 2016).

20 A well-recognized area in which non-B DNA has been linked to human genetic disease is that of
21 microsatellite repeat diseases (MRDs). MRDs represent a class of pathological conditions, currently
22 numbering >30, caused by the expansion of tandem repeats, mostly trinucleotide repeats, within
23 human genes (Table 2; Brouwer et al., 2009; Lopez Castel et al., 2010). In most cases, slipped/hairpin
24 DNA, G4 and triplex structures is believed to drive the expansion process (Lopez Castel et al., 2010;
25 Wells and Ashizawa, 2006). MRDs may be broadly classified into two types: type 1, in which
26 expansions are large (usually hundreds of copies of the repeat) and occur in non protein-coding
27 regions (untranslated and intronic) and type 2, in which expansions are less severe but occur in
28 protein-coding regions thereby altering protein sequence.

29 <TABLE 2 NEAR HERE>

30 Type 1 MRDs

31 In fragile X syndrome (FXS), a CGG repeat is expanded in the 5' UTR of the FMR1 gene. In Friedreich
32 ataxia (FA), expansions of a GAA repeat occur in the first intron of the frataxin (FRDA) gene, whereas
33 myotonic dystrophy (DM) is caused by the expansion of either a CTG repeat in the 3' UTR of the
34 DMPK gene (myotonic dystrophy type 1, DM1) or a CCTG repeat in the first intron of the ZNF9 gene
35 (myotonic dystrophy type 2, DM2). In amyotrophic lateral sclerosis with frontotemporal lobar
36 degeneration (ALS-FTLD) and spinocerebellar ataxia 36 (SCA36), expansions involve the G4-forming
37 GGGCC and GGCCTG exanucleotide repeats in the first intron of the C9orf72 and NOP56 genes,
38 respectively (Table 2).

39 In FXS and FA patients, repeat expansion is associated with histone markers specific for
40 heterochromatin (i.e. nontranscribed DNA), including deacetylation of histones H3 and H4 and
41 methylation of histone H3 at lysine 9 (H3K9me), supporting the notion that loss-of-function due to
42 gene silencing is the likely mechanism responsible for these diseases. In the case of DM, most of the
43 pathogenesis has been recapitulated in mouse models by experiments that show RNA gain-of-

1 function and the altered activity of at least two proteins, MBLN1 (muscleblind-like 1) and CUGBP1
2 (CUG-binding protein 1) (Figure 3).

3 <FIGURE 3 NEAR HERE>

4 MBNL1 is a key regulator of alternative splicing that binds to double-stranded RNA hairpins,
5 including those formed by CUG repeats (Lee and Cooper, 2009). In cells from DM1 patients, most
6 MBNL1 is found in discrete nuclear foci where it is sequestered by expanded CUG (or CCUG)
7 hairpins. Several transcripts are aberrantly spliced in DM1 mouse models, including *Clcn1* (chloride
8 channel 1), *Insr* (insulin receptor) and *Tnnt2* (cardiac troponin T). Thus, loss-of-function of MBNL1
9 mediated by long RNA hairpins is thought to contribute to DM pathology (Figure 3a). The second
10 protein affected, CUGBP1, binds single-stranded CUG repeats and becomes hyperphosphorylated by
11 PKC (protein kinase C) on CUG/CCUG repeat expansion. Activation of CUGBP1 exacerbates the
12 splicing defects of MBNL1 deficiency. It also increases the translation of embryonic transcripts, such
13 as *MEF2A*, and prevents the decay of short-lived mRNAs, including *c-FOS* and *TNF α* (tumour necrosis
14 factor alpha). The antagonistic activities of MBNL1 and CUGBP1 are critical for the shifts in
15 alternative splicing that accompany the transition from the embryonic to the adult developmental
16 stages; DM may therefore entail a reversal of this developmental pattern (Figure 3b). RNA gain-of-
17 function is also thought to underlie the pathology of fragile X-associated tremor/ataxia syndrome
18 (FXTAS), affecting older patients carrying moderate CGG expansions in the frataxin gene, as well as
19 *SCA8*, *SCA10* and *SCA12* (Brouwer et al., 2009). FMRP, whose levels are strongly reduced in FXS, is a
20 G4- and RNA-binding protein that inhibits translation of bound mRNA partners; in neurons, ~400
21 transcripts have been reported as potential FMRP binding targets. FMRP is also synthesized in situ
22 in dendrites, where synaptic activation induces its dephosphorylation, which in turns releases the
23 bound mRNAs and restores translation (Simone et al., 2015). Mounting evidence also supports a role
24 for G4 and other RNA secondary structures in mRNA transport along neurites for local protein
25 synthesis through their binding by FMRP and other proteins including TDP-43, FUS/TLS, hnRNPs, and
26 ZBP1 (Ishiguro et al., 2016)). A gain-of-function pathology has been attributed to G4 structures in
27 expanded *C9orf72* mRNAs, which by sequestering nucleolin and RNA-processing factors compromise
28 ribosomal RNA biogenesis, RNA editing and splicing (Figure 4). A shared mechanism through which
29 expanded repeats elicit pathological consequences in *SCA8*, DM1, FXTAS and FTD-ALS is their
30 induction of “Repeat-associated non-ATG translation” (RAN), whereby mRNA (both sense and
31 antisense) translation in all three reading frames occurs within the repeats themselves, yielding toxic
32 dipeptides (Figure 4). Lastly, transcription through expanded repeats may lead to the accumulation
33 of R-loops (persistent RNA-DNA hybrids between the nascent RNA and DNA template), which causes
34 abortive transcription and the induction of a DNA damage response.

35 Type 2 MRDs

36 In type 2 diseases, triplet repeat expansions occur exclusively in coding exons and hence they alter
37 protein sequence by increasing the length of homopolymeric amino acid runs, typically
38 polyglutamine and polyalanine (Table 2; Brouwer et al., 2009; Lopez Castel et al., 2010).
39 Homopolymeric amino acid runs contribute to ‘protein disorder’, a structural property that plays a
40 critical role in mediating protein–protein and protein–DNA interactions, particularly in protein
41 families comprising transcription factors and regulators of transcription and DNA replication.
42 Expansion beyond a critical threshold destabilises protein structure, causing aggregation or loss of
43 the binding affinities with partner molecules.

1 The number of tandem repeats at a given locus is often polymorphic in the general population
2 (variable number of tandem repeats or VNTR), a process that may be driven by slipped hairpins or
3 aberrant DNA replication. The presence of a VNTR within an intron or the regulatory region of a gene
4 is often associated with a change in its transcription, thereby yielding functionally different alleles
5 that may contribute to variable phenotypic traits (e.g. blood pressure, heart rate and muscular
6 tension) and susceptibility to multigenic diseases, such as brain disorders, obesity and stroke
7 (Hannan, 2010). Thus, slipped DNA may also contribute to phenotypic variability and susceptibility to
8 disease.

9 Non-B DNA/RNA and Cancer

10 Non-B DNA structures have been suggested to contribute to chromosomal rearrangements
11 (translocations and deletions) in cancer genomes. An analysis of ~20,000 translocation and ~46,000
12 deletion breakpoints revealed that the chance of finding a non-B DNA-forming repeat within ± 500
13 bp of these sites before any rearrangement occurred was greater than expected by chance, and that
14 the number of non-B DNA-forming repeats peaked exactly at breakpoints. These results were
15 interpreted to mean that rearrangements were promoted by the formation of non-B DNA structures,
16 possibly following their recognition and cleavage by DNA repair proteins or other nucleases (Bacolla
17 et al., 2016; Figure 5). Array CGH also indicated an overrepresentation of non-B DNA-forming
18 repeats in the proximity of DNA breaks of somatic copy number alterations in osteosarcoma (Smida
19 et al., 2017).

20 Activation of translation has emerged as a potent inducer of oncogenic transformation. A key factor
21 in the activation process, which is required for leukaemia maintenance, is eIF4A, an RNA helicase
22 that recognizes a subset of genes, including transcription factors and oncogenes, through RNA G4
23 structures in 5' UTRs. If not resolved these structures inhibit translation, and inhibition of translation
24 by silvestrol and other agents that target eIF4A reduced the expression levels of MYC, MYB, NOTCH,
25 CDK6, BCL2 and other oncogenes (Wolfe et al., 2014). The relevance of RNA G4 structures in
26 blocking translation was also highlighted by the fact that mice knockout for DHX36, which encodes
27 the RNA G4 resolvase RHAU, was embryonic lethal due to severe heart defects (Nie et al., 2015).
28 Overall, an increasing number of non-B DNA/RNA structure resolvases are being reported, which
29 unwind DNA and RNA secondary structures and enable the progression of DNA replication, RNA
30 transcription and translation complexes, and whose mutations are associated with human disease
31 and cancer. Examples include ATRX, BLM, CHL1, CSA, CSB, DHX9, DNA2, FANCI, PIF1, RTEL1, XPB,
32 XPD, XRCC1-XPF and WRN (Barthelemy et al., 2016; Gray et al., 2014; Jain et al., 2013; Lu et al.,
33 2015; Maizels, 2015; Paeschke et al., 2013; Scheibye-Knudsen et al., 2016). As mentioned, lncRNAs
34 are emerging as powerful repressors of gene expression in part by forming DNA:RNA triplexes, and
35 by recruiting chromatin repressive complexes, such as the polycomb repressive complex 2 (PRC2)
36 and modified histones (H3K27me3 and H3K9me3) (Mondal et al., 2015; Grote and Herrmann, 2013).
37 From 17 gene expression studies representing 2,999 primary breast tumors, the lncRNA MEG3 levels
38 were on average the lowest, and MEG3 expressed at low levels in high-grade breast tumours
39 (Mondal et al., 2015), suggesting that insufficient gene repression by lncRNA-mediated RNA:DNA
40 triplexes contributes to tumorigenesis.

41 In summary, the formation of DNA and RNA secondary structures contribute to a number of quite
42 distinct human pathologies, where mutagenesis is elicited in various ways: DNA breakage and
43 stimulation of recombination leading to genomic rearrangements, altered RNA secondary structure,
44 mRNA codon changes resulting in impaired protein function, and altered transcription and
45 translation affecting protein levels. See also Causes and Consequences of Structural Genomic

1 Alterations in the Human Genome; Chromosomal Genetic Disease: Structural Aberrations, and
2 Segmental Duplications and Genetic Disease; Identifying Genes Underlying Human Inherited Disease;
3 Microsatellites; Mechanisms of RNA-Induced Toxicity in Diseases Characterised by CAG Repeat
4 Expansions; Trinucleotide Repeat Expansions: Disorders; Gene Structure and Organization; DNA
5 Helicases; DNA Helicase-deficiency Disorders; The Transcription/DNA Repair Factor TFIIH; mRNA
6 Untranslated Regions (UTRs); Protein Synthesis in Neurons.

7

8 Mechanisms Underlying DNA- and RNA-structure Induced Diseases

9 A large number of studies have been conducted in model organisms, including bacteria, fly, yeast,
10 mammalian cell culture and mice to address the mechanisms responsible for non-B DNA-induced
11 genetic instability, particularly in the context of repeat expansion related to MRDs. Knocking down
12 the expression of proteins involved in DNA replication, repair and recombination affected repeat
13 stability by either increasing or decreasing instability (depending on the specific enzyme), suggesting
14 that these biochemical processes are all potentially involved (Lopez Castel et al., 2010; Wells and
15 Ashizawa, 2006; Polleys et al., 2017). Despite these important advances, the mechanisms restricting
16 the timing of repeat instability and genomic rearrangements to early development remain elusive. In
17 transgenic mice, the presence of Msh2 and Msh3, two proteins of the mismatch repair pathway, was
18 found to be critical for eliciting microsatellite repeat expansion during all stages of development
19 (Zhao et al., 2015). A role restricted to adult somatic tissues has been noted for ataxia telangiectasia
20 and Rad3 related (Atr), post-meiotic segregation increased 2 (Pms2) and the Ogg1 glycosylase. Atr,
21 Pms2 and Ogg1 are involved in DNA damage responses and DNA repair. Thus, DNA damage and
22 repair may play a critical role in eliciting non-B DNA-induced genetic instability, at least in the
23 context of MRDs. Concerning genomic rearrangements, classic and alternative nonhomologous end-
24 joining, DSB repair pathways that can be error-prone are believed to act on DSBs induced at sites of
25 non-B DNA structures.

26 Numerous studies support an intimate connection between unresolved non-B DNA structures and
27 genomic instability. For example, TOP1, which releases negative supercoiling (negative supercoiling
28 promotes non-B DNA), is recruited to chromatin by BRG1 (product of the SMARCA4 gene), a
29 component of the SWI/SNF chromatin remodeler complex, and FACT, a two-protein complex
30 associated directly with histones. siRNA knock down of either TOP1 or SMARCA4 increased
31 translocation frequencies dramatically in B-cells ??? (Husain et al., 2016). In the context of CS, the
32 accumulation of non-B structures in mitochondrial DNA appears to stall rDNA transcription and to
33 lead to persistent activation of a DNA damage response which, orchestrated by PARP1, results in
34 NAD⁺ depletion and increased lactate production (Scheibye-Knudsen et al., 2016). Unresolved
35 CTG/CAG hairpin structures in FANCD1 deficient cells led to loss of PCR amplifiable genomic
36 fragments, likely due to gross rearrangements, at loci associated with developmental disorders and
37 cancer (Barthelemy et al., 2016). Fancj-null mice, in addition to displaying increased susceptibility to
38 epithelial tumors and intrastrand crosslinking agents, also exhibited increased microsatellite
39 instability (Matsuzaki et al., 2015). These composite observations support a role for FANCD1 in
40 resolving non-B DNA structures during DNA replication and the involvement of its mutations in
41 human cancer.

42 See also Trinucleotide Repeat Expansions: Mechanisms and Disease Associations; Mechanisms of
43 Chromosome Translocations in Cancer; Genomic Rearrangements: Mutational Mechanisms.

44 Prospects for the Future

1 The relevance of unusual or aberrant DNA structures to human disease has extended the interest in
2 this area from the narrow confines of the nucleic acid biochemistry/physicochemical laboratory to a
3 wider community operating in medicine, clinical diagnostics, genetics and bioinformatics, thereby
4 propelling the field into mainstream biomedical research. Although the transient nature of DNA
5 secondary structures remains a challenge, the development of non-B DNA structure-specific
6 antibodies and CHIP-seq technology is expected to better define the landscape of non-B DNA
7 structures in living human cells.

8 The advent of aCGH is likely to reveal more instances of the co-localisation of repetitive DNA with
9 CNVs, both in the context of phenotypic variation, as well as in the sphere of susceptibility to
10 infectious disease, cancer and inherited disease. As more information is acquired from mouse
11 models on the timing of repeat instability during development and its relationships with chromatin
12 remodelling, CpG methylation-linked epigenetic reprogramming, and DNA damage and repair,
13 experiments can be designed to further address the biochemical pathways involved in non-B-
14 directed genetic instability. Strong advances have been made that highlighted biological roles for
15 non-B DNA/RNA structures, and future studies are expected that will probe the spectrum of non-B
16 DNA and RNA biological roles. Likewise, the number of enzymes interacting with non-B DNA and
17 RNA has increased considerably, which is revealing more intimate connections between non-B
18 structures and human pathology. Sequencing cancer genomes was crucial to address the question as
19 to whether non-B DNA is responsible for at least some of the genomic rearrangements associated
20 with cancer. As larger cancer genome datasets will be available, it will be possible to assess the
21 quantitative contribution of non-B DNA to mutational loads in cancer.

22 Acknowledgements

23 This work was supported by the National Institutes of Health Grant CA92584 to J.A.T. Work by the
24 authors used the Extreme Science and Engineering Discovery Environment (XSEDE), which is
25 supported by National Science Foundation grant number ACI-1548562, and the Texas Advanced
26 Computing Center, supported by the National Science Foundation grant number ACI-1134872.

27 Glossary

28 aCGH

29 Array comparative genomic hybridisation (a technique used to detect genomic copy number
30 variations at a high resolution).

31 CNV

32 Copy number variant (segment of DNA that exhibits copy-number differences when two or more
33 genomes are compared).

34 Microsatellites

35 Simple sequence repeats that contain sequences of 1–6 base pairs of DNA.

36 Non-B DNA

37 DNA structures, distinct from the canonical B-form, including left-handed Z-DNA, cruciforms, looped-
38 out or slipped folds, parallel DNA, triplexes and quadruplexes.

39 VNTR

1 A location in the genome where a short nucleotide sequence is organised as a tandem repeat which
2 displays variations in length between individuals.

3 References

- 4 Bacolla A, Tainer JA, Vasquez KM and Cooper DN (2016) Translocation and deletion breakpoints in
5 cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids
6 Research* 44: 5673-5688.
- 7 Bacolla A and Wells RD (2004) Non-B DNA conformations, genomic rearrangements, and human
8 disease. *Journal of Biological Chemistry* 279: 47411–47414.
- 9 Barthelemy J, Hanenberg H and Leffak M (2016) FANCI is essential to maintain microsatellite
10 structure genome-wide during replication stress. *Nucleic Acids Research* 44: 6803-6816.
- 11 Bena F, Gimelli S, Migliavacca E et al. (2010) A recurrent 14q32.2 microdeletion mediated by
12 expanded TGG repeats. *Human Molecular Genetics* 19: 1967–1973.
- 13 Bonaglia MC, Giorda R, Massagli A et al. (2009) A familial inverted duplication/deletion of 2p25.1-
14 25.3 provides new clues on the genesis of inverted duplications. *European Journal of Human
15 Genetics* 17: 179–186.
- 16 Brouwer JR, Willemsen R and Oostra BA (2009) Microsatellite repeat instability and neurological
17 disease. *BioEssays* 31: 71–83.
- 18 Chen X, Shen Y, Zhang F et al. (2013) Molecular analysis of a deletion hotspot in the NRXN1 region
19 reveals the involvement of short inverted repeats in deletion CNVs. *The American Journal of Human
20 Genetics* 92: 375-386.
- 21 Chuzhanova N, Chen JM, Bacolla A et al. (2009) Gene conversion causing human inherited disease:
22 evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in
23 DNA breakage and repair. *Human Mutation* 30: 1189–1198.
- 24 Conrad DF, Pinto D, Redon R et al. (2010) Origins and functional impact of copy number variation in
25 the human genome. *Nature* 464: 704–712.
- 26 Dong DW, Pereira , Barrett SP et al. (2014) Association of G-quadruplex forming sequences with
27 human mtDNA deletion breakpoints. *BMC Genomics* 15: 677.
- 28 Francois M, Leifert WR, Hecker J, Faunt J and Fenech MF (2016) Guanine-quadruplexes are increased
29 in mild cognitive impairment and correlate with cognitive function and chromosomal DNA
30 damage. *DNA Repair* 46: 29-36.
- 31 Gray LT, Vallur AC, Eddy J and Maizels N (2014) G quadruplexes are genomewide targets of
32 transcriptional helicases XPB and XPD. *Nature Chemical Biology* 10: 313-318.
- 33 Grote P and Herrmann BG (2013) The long non-coding RNA Fendrr links epigenetic control
34 mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biology* 10: 1579-
35 1585.
- 36 Hannan AJ (2010) Tandem repeat polymorphisms: modulators of disease susceptibility and
37 candidates for ‘missing heritability’. *Trends in Genetics* 26: 59–65.

- 1 Hansel-Hertsch R, Beraldi D, Lensing SV et al. (2016) G-quadruplex structures mark human regulatory
2 chromatin. *Nature Genetics* 48: 1267-1272.
- 3 Husain A, Begum NA, Taniguchi T et al. (2016) Chromatin remodeler SMARCA4 recruits
4 topoisomerase I and suppresses transcription-associated genomic instability. *Nature*
5 *Communications* 7: 10549.
- 6 Ishiguro A, Kimura N, Watanabe Y, Watanabe S and Ishihama A (2016) TDP-43 binds and transports
7 G-quadruplex-containing mRNAs into neurites for local translation. *Genes to Cells* 21: 466-481.
- 8 Jain A, Bacolla A, Del Mundo IM et al. (2013) DHX9 is involved in preventing genomic instability
9 induced by alternative structured DNA in human cells. *Nucleic Acids Research* 41: 10345-10357.
- 10 Kamat MA, Bacolla A, Cooper D and Chuzhanova N (2015) A role for non-B DNA forming sequences
11 in mediating microlesions causing human inherited disease. *Human Mutation* 37: 65-73.
- 12 Kornreich R, Bishop DF and Desnick RJ (1990) Alpha-galactosidase A gene rearrangements causing
13 Fabry disease. Identification of short direct repeats at breakpoints in an Alu-rich gene. *Journal of*
14 *Biological Chemistry* 265: 9319–9326.
- 15 Kurahashi H, Inagaki H, Ohye T et al. (2010) The constitutional t(11;22): implications for a novel
16 mechanism responsible for gross chromosomal rearrangements. *Clinical Genetics*. [Epub ahead of
17 print; doi: 10.1111/j.1399-0004.2010.01445.x].
- 18 Lange J, Skaletsky H, van Daalen SK et al. (2009) Isodicentric Y chromosomes and sex disorders as
19 byproducts of homologous recombination that maintains palindromes. *Cell* 138: 855–869.
- 20 Lee JE and Cooper TA (2009) Pathogenic mechanisms of myotonic dystrophy. *Biochemical Society*
21 *Transactions* 37: 1281–1286.
- 22 Liu HY, Zhao Q, Zhang TP et al. (2016) Conformation selective antibody enables genome profiling and
23 leads to discovery of parallel G-quadruplex in human telomeres. *Cell Chemical Biology* 23: 1261-
24 1270.
- 25 Lopez Castel A, Cleary JD and Pearson CE (2010) Repeat instability as the basis for human diseases
26 and as a potential target for therapy. *Nature Reviews. Molecular Cell Biology* 11: 165–170.
- 27 Lu S, Wang G, Bacolla A et al. (2015) Short inverted repeats are hotspots for genetic instability:
28 Relevance to cancer genomes. *Cell Reports* 10: 1674-1680.
- 29 Maizels N (2015) G4-associated human diseases. *EMBO Reports* 16: 910-922.
- 30 Matsuzaki K, Borel V, Adelman CA, Schindler D and Boulton SJ (2015) FANCI suppresses
31 microsatellite instability and lymphomagenesis independent of the Fanconi anemia pathway. *Genes*
32 *& Development* 29: 2532-2546.
- 33 Mondal T, Subhash S, Vaid R et al. (2015) MEG3 long noncoding RNA regulates the TGF- β pathway
34 genes through formation of RNA-DNA triplex structures. *Nature Communications* 6: 7743.
- 35 Neidle S (2009) The structures of quadruplex nucleic acids and their drug complexes. *Current*
36 *Opinion in Structural Biology* 19: 239–250.
- 37 Nie J, Jiang M, Zhang X et al. (2015) Post-transcriptional regulation of Nkx2-5 by RHAU in heart
38 development. *Cell Reports* 13, 723-732, 2015.

- 1 Paeschke K, Bochman ML, Garcia PD et al. (2013) Pif1 family helicases suppress genome instability at
2 G-quadruplex motifs. *Nature* 497: 458-462.
- 3 Perry GH, Ben-Dor A, Tsalenko A et al. (2008) The fine-scale and complex architecture of human
4 copy-number variation. *American Journal of Human Genetics* 82: 685–695.
- 5 Polleys EJ, House NCM and Freudenreich CH (2017) Role of recombination and replication fork
6 restart in repeat instability. *DNA Repair* 56: 156-165.
- 7 Quemener S, Chen JM, Chuzhanova N et al. (2010) Complete ascertainment of intragenic copy
8 number mutations (CNMs) in the CFTR gene and its implications for CNM formation at other
9 autosomal loci. *Human Mutation* 31: 421–428.
- 10 Quental R, Azevedo L, Rubio V, Diogo L and Amorim A (2009) Molecular mechanisms underlying
11 large genomic deletions in ornithine transcarbamylase (OTC) gene. *Clinical Genetics* 75: 457–464.
- 12 Rooms L, Reyniers E and Kooy RF (2007) Diverse chromosome breakage mechanisms underlie
13 subtelomeric rearrangements, a common cause of mental retardation. *Human Mutation* 28: 177–
14 182.
- 15 Scheibye-Knudsen M, Tseng A, Jensen MB et al. (2016) Cockayne syndrome group A and B proteins
16 converge on transcription-linked resolution of non-B DNA. *Proceedings of the National Academy of
17 Sciences of the USA* 113: 12502-12507.
- 18 Shin S, Ham S, Park J et al. (2016) Z-DNA-forming sites identified by ChIP-Seq are associated with
19 actively transcribed regions in the human genome. *DNA Research* 23: 477-486.
- 20 Simone R, Fratta P, Neidle S, Parkinson GN and Isaacs AM (2015) G-quadruplexes: Emerging roles in
21 neurodegenerative diseases and the non-coding transcriptome. *FEBS Letters* 589: 1653-1668.
- 22 Skaletsky H, Kuroda-Kawaguchi T, Minx PJ et al. (2003) The male-specific region of the human Y
23 chromosome is a mosaic of discrete sequence classes. *Nature* 423: 825–837.
- 24 Smida J, Xu H, Zhang Y et al. (2017) Genome-wide analysis of somatic copy number alterations and
25 chromosomal breakages in osteosarcoma. *International Journal of Cancer* 141: 816-828.
- 26 Verdin H, D'haene B, Beysen D et al. (2013) Microhomology-mediated mechanisms underlie non-
27 recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain. *PLOS Genetics*
28 9: e1003358.
- 29 Vissers LE, Bhatt SS, Janssen IM et al. (2009) Rare pathogenic microdeletions and tandem
30 duplications are microhomology-mediated and stimulated by local genomic architecture. *Human
31 Molecular Genetics* 18: 3579–3593.
- 32 Wells RD and Ashizawa T (2006) *Genetic Instabilities and Neurological Diseases*, 2nd edn. San Diego,
33 CA: Elsevier/Academic Press.
- 34 Wolfe AL, Singh K, Zhong Y et al. (2014) RNA G-quadruplexes cause eIF4A-dependent oncogene
35 translation in cancer. *Nature* 513: 65-70.
- 36 Zhang F, Seeman P, Liu P et al. (2010) Mechanisms for nonrecurrent genomic rearrangements
37 associated with CMT1A or HNPP: rare CNVs as a cause for missing heritability. *American Journal of
38 Human Genetics* 86: 892–903.

1 Zhao J, Bacolla A, Wang G and Vasquez KM (2010) Non-B DNA structure-induced genetic instability
2 and evolution. Cellular and Molecular Life Sciences 67: 43–62.

3 Zhao XN, Kumari D, Gupta S et al. (2015) Muts β generates both expansions and contractions in a
4 mouse model of the Fragile X-associated disorders. Human Molecular Genetics 24: 7087-7096.

5

6 Further Reading

7 Avino A, Mazzini S, Gargallo R and Eritja R (2016) The effect of small cosolutes that mimic molecular
8 crowding conditions on the stability of triplexes involving duplex DNA. International Journal of
9 Molecular Sciences 17: 211.

10 Bacolla A, Wang G and Vasquez KM (2015) New perspectives on DNA and RNA triplexes as effectors
11 of biological activity. PLOS Genetics 11: e1005696.

12 Choi J and Majima T (2011) Conformational changes of non-B DNA. Chemical Society Reviews 40:
13 5893-5909.

14 D'Angelo CS, Gajecka M, Kim CA et al. (2009) Further delineation of nonhomologous-based
15 recombination and evidence for subtelomeric segmental duplications in 1p36 rearrangements.
16 Human Genetics 125: 551–563.

17 Karlin S, Brocchieri L, Bergman A, Mrazek J and Gentles AJ (2002) Amino acid runs in eukaryotic
18 proteomes and disease associations. Proceedings of the National Academy of Sciences of the USA
19 99: 333–338.

20 Orr HT and Zoghbi HY (2007) Trinucleotide repeat disorders. Annual Review of Neuroscience 30:
21 575–621.

22 [Repping S, Skaletsky H, Lange J et al. \(2002\) Recombination between palindromes P5 and P1 on the](#)
23 [human Y chromosome causes massive deletions and spermatogenic failure. American Journal of](#)
24 [Human Genetics 71: 906–922.](#)

25 Schofield JPR, Cowan JL and Coldwell MJ (2015) G-quadruplexes mediate local translation in
26 neurons. Biochemical Society Transactions 43, 338-342.

27 Stankiewicz P and Lupski JR (2010) Structural variation in the human genome and its role in disease.
28 Annual Review of Medicine 61: 437–455.

29 Wang G and Vasquez KM (2009) Models for chromosomal replication-independent non-B DNA
30 structure-induced genetic instability. Molecular Carcinogenesis 48: 286–298.

31

32 Cross-references to other eLS articles: As eLS is continually growing, authors writing update articles
33 should check the validity of any cross references and add new cross references as appropriate. This
34 is an intrinsic strength of the eLS format. In order to create a cross-reference please cite the other
35 article's unique DOI and place it at the end of the relevant paragraph. As shown in the following
36 example:

37 “The reversal potential defines the potential at which the net movement of charged ions into and
38 out of the channel is zero, and for glutamate-activated channels is approximately 0 mV, whereas

1 that for GABA-activated channels it is typically around -70 mV. See also: DOI:
2 10.1002/9780470015902.a0005907”

3 To identify appropriate eLS articles to cross-reference, go to www.els.net where abstracts, key
4 concepts, keywords and figures for eLS articles are open to all browsers for quick reference. You can
5 browse by topic and narrow by sub-topic, or enter text searches to track down related articles.

6 To find an article’s DOI and build it in to your manuscript as a cross-reference, click on the article
7 title. On the article landing page, the DOI is listed under the title and author information in the top
8 left of the screen.

9 References and Further Reading: eLS articles may contain up to 50 cited References and up to 10
10 uncited Further Reading items. Both sections must be present. If important new papers have been
11 published, they should be cited at relevant locations in the article text and listed in the References
12 section. If textbooks, reference works, etc. were included in the Further Reading section, they should
13 be updated with the latest editions.

14

15 Glossary: Inclusion of a Glossary is advised for all scientific articles. Please consider including a list of
16 6-10 terms used in the article that require definition for the target audience, assuming Advanced
17 article readers have a greater level of knowledge. Your list may be edited by the publishers to ensure
18 consistency, and to remove unnecessary overlap.

19

20

21 FIGURES/TABLES: In your submission e-mail, please indicate those figures and tables that have been
22 altered and those, including legends, that remain entirely unchanged from the previous version. We
23 can then alert our typesetters accordingly.

24

25 To facilitate peer review, all NEW OR MODIFIED FIGURES AND TABLES, should be placed at the end
26 of the manuscript file. Please do not embed them in the body text of your article. Our typesetters
27 will take responsibility for inserting each figure or table as close as possible to your text citation (e.g.
28 “See Figure 1..” ..”as seen in Table 2”). Please ensure that you cite all figures and tables in the text.

29 NEW OR MODIFIED FIGURES (though not tables) should also be supplied separately in high
30 resolution, editable figure files. We recommend EPS files for line illustrations. PDF, PPT, DOC, AI,
31 CDR, WMF and INDD files can also be used; provided they are in vector format (i.e. still editable).
32 TIFF files should only be used for unlabelled, unedited photographs.

33

34 Line illustrations supplied electronically should not have a resolution less than 600 dpi. Tone
35 illustrations (black and white and colour photographs for instance) should have a resolution of
36 around 300 dpi. Please do NOT supply ‘thumbnails’ or low resolution images).

37

38 Figure and Table Permissions: We must remind you that it is the author’s responsibility to obtain
39 permission to reproduce any figures or tables that have been published elsewhere (note this only

1 applies to new figures or tables, you do not need to seek permission to re-use figures or tables from
2 the current version of this eLS article).

3

4 Please note that even if you have modified a copyrighted figure or table, it is still likely to require
5 permission.

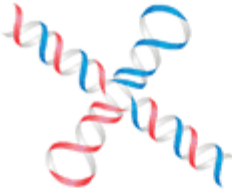
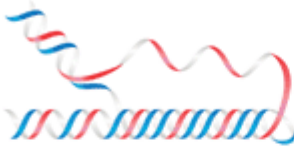
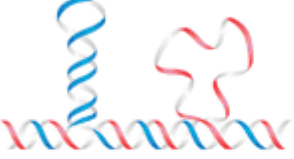
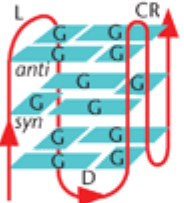
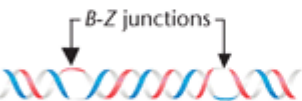
6

7 To help with the applications process, we have provided you with a permissions request form,
8 attached to your Author Guidelines email.

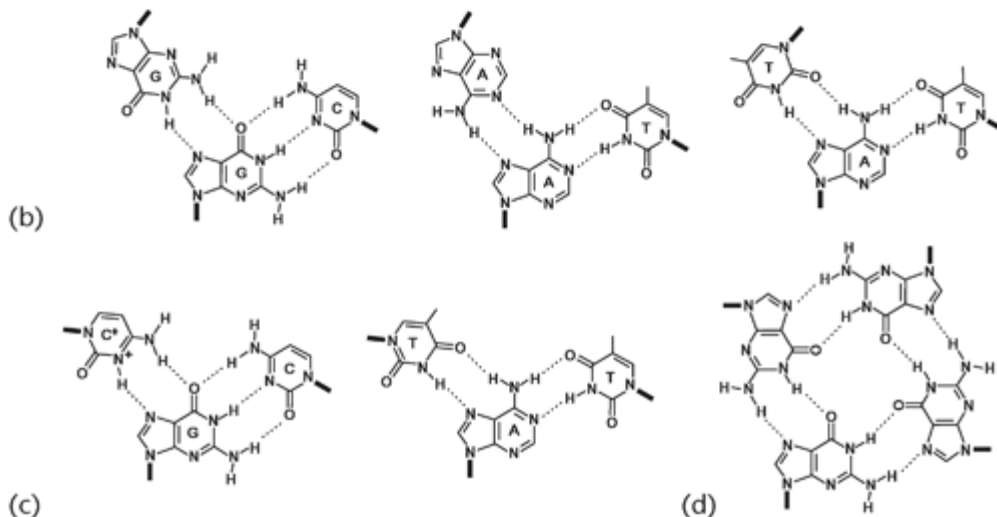
9

10 Please confirm at the bottom of your manuscript the numbers of the figures or tables that do not
11 require permission.

12

Name	Conformation	General seq. requirements	Sequence
Cruciform		Inverted repeats	$\begin{array}{c} \overrightarrow{\text{TCGGT}_x\text{ACCGA}} \\ \overleftarrow{\text{AGCCAyTGGCT}} \end{array}$
Triplex		$(R \cdot Y)_n$ mirror repeats	$\begin{array}{c} \overrightarrow{\text{AAGAGG}_x\text{GGAGAA}} \\ \overleftarrow{\text{TTCTCC}_y\text{CCTCTT}} \end{array}$
Slipped (hairpin) structure		Direct repeats	$\begin{array}{c} \overrightarrow{\text{TCGGTTCGGT}} \\ \overrightarrow{\text{AGCCAAGCCA}} \end{array}$
Quadruplex		Oligo $(G)_n$ tracts	$\text{AG}_3(\text{T}_2\text{AG}_3)_3$ Single strand
Left-handed Z-DNA		$(YR \cdot YR)_n$	$\begin{array}{c} \text{CGCGTGC GTGTG} \\ \text{GCGCACGCACAC} \end{array}$

(a)



1

2 Figure 1. Non-B DNA structures formed by genomic repetitive sequences. (a) Most common non-B
3 DNA conformations, ribbon models of helical foldings, repetitive motifs requirement and example of
4 sequences. Center dot, Watson–Crick hydrogen bond interactions; x,y, nucleotides in the spacer
5 between repeats; L, lateral loop; D, diagonal loop and CR, chain reversal loop. For cruciform DNA, an
6 extended conformation is shown. For triplex DNA, a 3' RRY isomer is depicted in which the 3'-half of
7 the purine-rich strand folds back to form the Hoogsteen-bound third strand. For quadruplex DNA, an

1 idealised structure is drawn to highlight the loop characteristics and the relative orientation of the
 2 syn and anti N-glycosidic configurations. (b) RRY base triplets showing the Hoogsteen-bound base
 3 (left). Thymine can be incorporated into RRY triplexes due to the symmetry of the carbonyl groups.
 4 (c) YRY base triplets showing the Hoogsteen-bound pyrimidine and the stabilisation afforded by
 5 cytosine protonation. (d) G-tetrad.

Table 1. Non-B DNA-forming repeats and human genes. Most relevant repetitive DNA sequences associated with human genes or gene classes

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	1450	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

6

Purine:pyrimidine tracts in introns of genes

Gene category/function	P-value	
	≥ 250 nt (228 genes)	≥ 100 nt (1951 genes)
Ion channel activity	1.95×10^{-05}	5.92×10^{-09}

Purine:pyrimidine tracts in introns of genes

Gene category/function	P-value	
Protein binding	3.14×10^{-03}	6.25×10^{-15}
Glutamate receptor activity	6.11×10^{-04}	1.92×10^{-07}
Cell adhesion	1.11×10^{-04}	3.36×10^{-12}
Cell communication	2.19×10^{-04}	5.24×10^{-15}
Transmission of nerve impulse	1.83×10^{-04}	5.24×10^{-08}
Synapse	2.18×10^{-02}	7.69×10^{-05}
Alternative splicing	ND	2×10^{-82}
Chromosomal translocations	ND	1×10^{-07}

1

Tetranucleotide repeats (TR) in introns of genes

Gene category/function/attribute	P-value	
Localisation to the membrane	1×10^{-07} – 5×10^{-30}	
Ion channel	5×10^{-02} – 1×10^{-13}	(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; 190–1423 genes/group)
Cell adhesion	8×10^{-04} – 2×10^{-37}	
Alternative splicing	1×10^{-64}	≥ 8 TR units (4182 genes)
Chromosomal translocations	2×10^{-07}	≥ 8 TR units (4182 genes)

1

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

Gene category/function	P-value (coding plus noncoding exons) (2626 genes)
Transcription regulator activity	2.0×10^{-40}
Regulation of cellular processes	2.3×10^{-38}
Protein binding	2.0×10^{-33}
Sequence-specific DNA binding	3.8×10^{-23}
Nuclear localisation	9.3×10^{-22}
RNA polymerase II transcription factor activity	1.2×10^{-16}
Axon guidance	2.3×10^{-05}
MAPK signalling pathway	2.1×10^{-04}
WNT signalling pathway	2.4×10^{-04}

2

 G-quadruplex in both 5'- and 3'-UTR

Gene category/function	P-value
------------------------	---------

Note: ND, not determined.

a

Source: With kind permission from Springer Science+Business Media (Zhao et al., 2010).

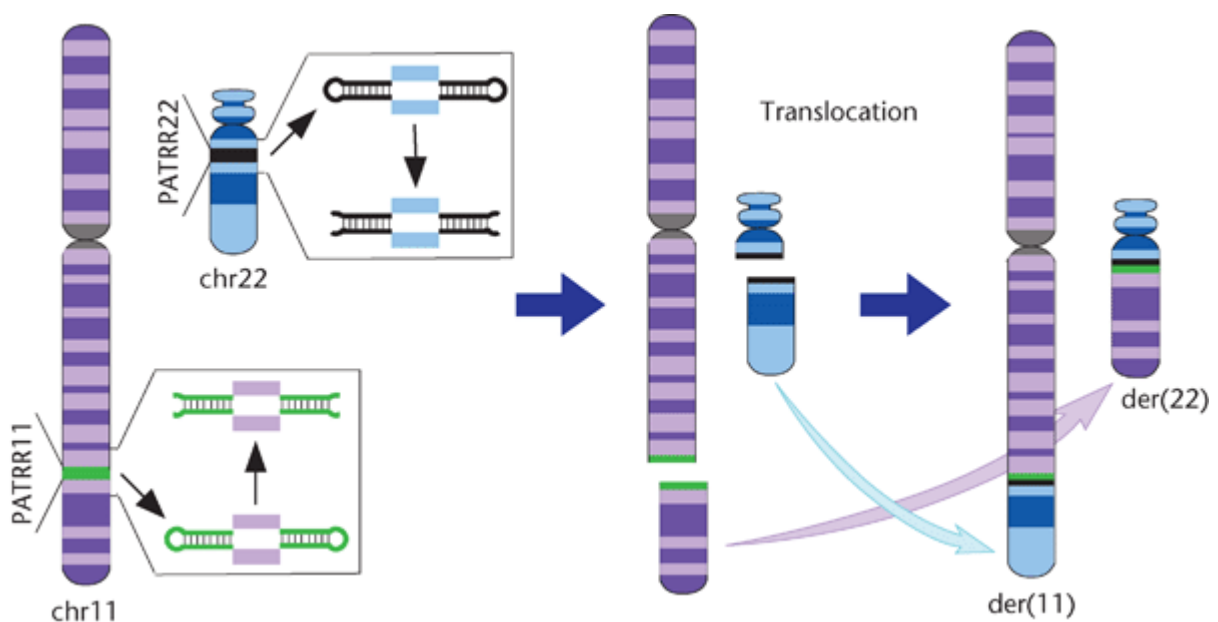
	5'-UTR	3'-UTR
Guanyl-nt exchange factor activity	7.9×10^{-13}	6.3×10^{-12}
Rho guanyl-nt exchange factor activity	7.9×10^{-10}	1.6×10^{-09}

G-quadruplex in both 5'- and 3'-UTR

Gene category/function	P-value	
Regulation of Rho signal transduction	2.0×10^{-10}	1.6×10^{-09}
Transcription factor activity	6.3×10^{-05}	3.2×10^{-10}
Sequence-specific DNA binding	2.5×10^{-06}	3.2×10^{-08}

1

2



3

4 Figure 2. Cruciform-mediated chromosomal t(11;22) translocation and the Emanuel syndrome. The
5 PATRR sequences on human chromosome 11 (green) and 22 (black) are proposed to fold into large
6 cruciform structures at some frequency during gametogenesis and be cleaved at the single-stranded
7 tips, resulting in double-strand breaks (left insets). The broken chromosomal ends (middle) join
8 aberrantly, yielding the derivative chromosomes der(11) and der(22) (right). Occasional inheritance
9 of der(22), in addition to a normal karyotype, is responsible for the Emanuel syndrome in the
10 offspring.

11

12

Table 2. Microsatellite repeat diseases (MRDs). Type 1 and type 2 microsatellite repeat diseases

Disease	Gene	Chromosome location	Amplet	Normal copy length or number	Expanded copy length or number	Location	Disease symbol
FRAXE-associated mental retardation	AFF2	Xq28	CCG	4–39	>200	5'-UTR	FRAXE
Spinocerebellar ataxia 10	ATXN10	22q13.31	ATTCT	<29	>800	Intron 1	SCA10
Spinocerebellar ataxia 8	ATXN8OS	13q21	CTG	15–50	110–130	3'-UTR	SCA8
Jacobsen syndrome	CBL2	11q23.3	CCG	11	700–800	?	FRA1B
Myotonic dystrophy type 2	CNBP	3q21	CCTG	104–176 bp	75–11 000	Intron 1	DM2
Epilepsy progressive myoclonic	CSTB	21q2.3	CCCCGCCCGCG	2–3	30–75	Promoter	EPM1
Mental retardation	DIP2B	12q13.13	CGG	6–23	250–285	5'-UTR	FRA12A
Myotonic dystrophy type 1	DMPK	19q13.32	CTG	<30	50–2000	3'-UTR	DM1

Source: With kind permission from Springer Science+Business Media (Zhao et al., 2010).

Type 1 diseases

Table 2. Microsatellite repeat diseases (MRDs). Type 1 and type 2 microsatellite repeat diseases

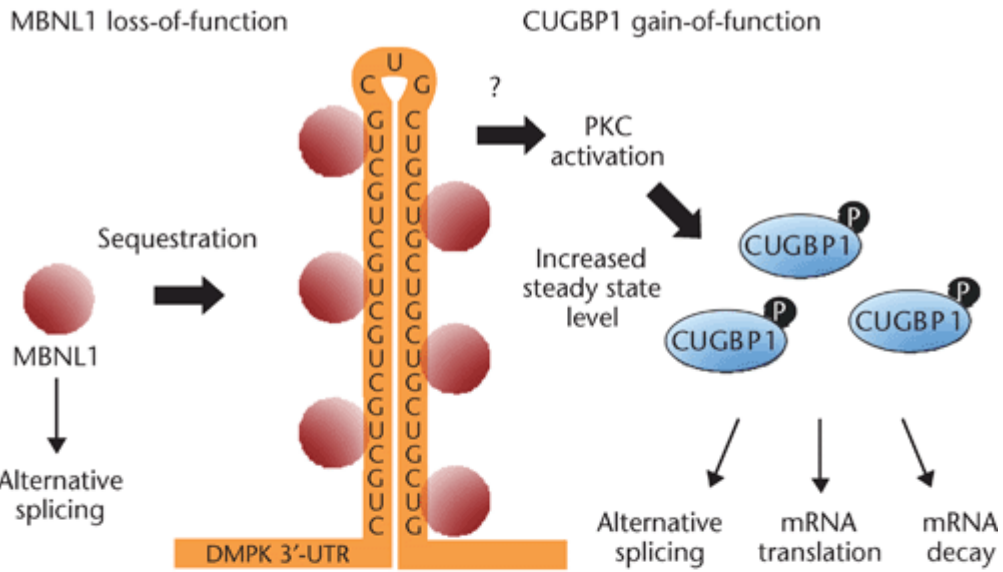
Disease	Gene	Chromosome location	Amplet	Normal copy length or number	Expanded copy length or number	Location	Disease symbol
Fragile X mental retardation syndrome	FMR1	Xq27.3	CGG	5–52	>200	5'-UTR	FXS
Fragile X-associated tremor ataxia syndrome	FMR1	Xq27.3	CGG	<55	55–200	5'-UTR	FXTAS
Friedreich ataxia	FXN	9q21.11	GAA	7–20	200–900	Intron 1	FA
Spinocerebellar ataxia 12	PPP2R2B	5q32	CAG	7–31	55–78	5'-UTR	SCA12
Cataract formation in myotonic dystrophy	SIX5	19q13.32	CTG	5–37	≥50	Promoter	SIX5
Spinocerebellar ataxia 31	TK2/BEAN	16q22	TGGAA	0	110	Intron	SCA31
Type 2 diseases							
Spino-bulbar muscular atrophy (Kennedy disease)	AR	Xq12	CAG	17–26	40–52	Coding region	SBMA

Table 2. Microsatellite repeat diseases (MRDs). Type 1 and type 2 microsatellite repeat diseases

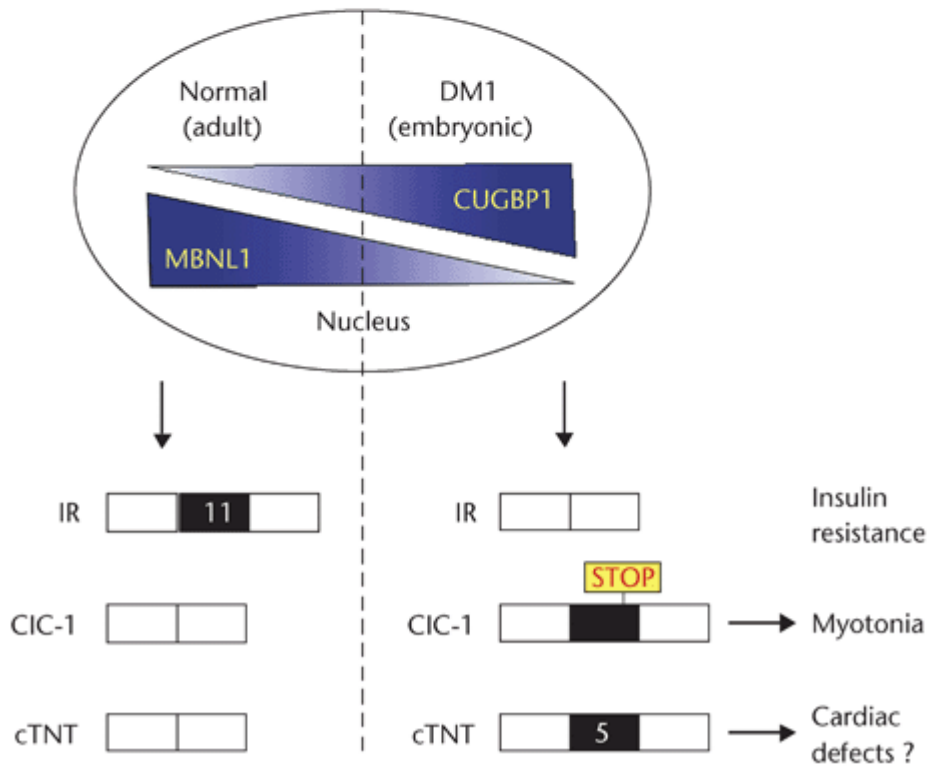
Disease	Gene	Chromosome location	Amplet	Normal copy length or number	Expanded copy length or number	Location	Disease symbol
Mental retardation and epilepsy	ARX	Xp21.3	GCG	10–16	17–33	Exon 1	XLMR
Dentatorubro-pallidoluysian atrophy (Haw river fever)	ATN1	12p13.31	CAG	3–36	48–93	Coding region	DRPLA
Spinocerebellar ataxia 1	ATXN1	6p22.3	CAG	6–44	37–91	Coding region	SCA1
Spinocerebellar ataxia 2	ATXN2	12q24.12	CAG	14–31	32–500	Coding region	SCA2
Machado–Joseph disease	ATXN3	14q32.12	CAG	13–47	53–86	Coding region	SCA3
Spinocerebellar ataxia 7	ATXN7	3p14.1	CAG	7–35	36–300	Coding region	SCA7
Spinocerebellar ataxia 6	CACNA1A	19p13.13	CAG	4–18	19–33	Coding region	SCA6
Blepharophimosis syndrome and premature ovarian failure 3	FOXL2	3q22.3	GCN	14	22–24	Coding region	BPES
Hand-foot-genital syndrome	HOXA13	7p15.2	GCN	14–18	22–30	Coding region	HFGS

Table 2. Microsatellite repeat diseases (MRDs). Type 1 and type 2 microsatellite repeat diseases

Disease	Gene	Chromosome location	Amplet	Normal copy length or number	Expanded copy length or number	Location	Disease symbol
Brachydactyly syndactyly syndrome	HOXD13	2q31.1	GCN	15	22–29	Coding region	SPD
Huntington disease	HTT	4p16.3–4p16.2	CAG	<35	40–400	Coding region	HD
Huntington disease-like 2	JPH3	16q24.2	CTG	6–27	44–57	Exon 1	HDL2
Oculopharyngeal muscular dystrophy	PABPN1	14q11.2	GCG	6	7–13	Coding region	OPMD
Congenital central hypoventilation syndrome	PHOX2B	4p13	GCN	20	25–33	Coding region	CCHS
Cleidocranial dysplasia	RUNX2	6p12.3	GCK	17	27	Coding region	CCD
Spinocerebellar ataxia 17	TBP	6q27	CAG	25–42	45–63	Coding region	SCA17
Holoprosencephaly	ZIC2	13q32.3	GCN	15	25	Coding region	HPA



(a)

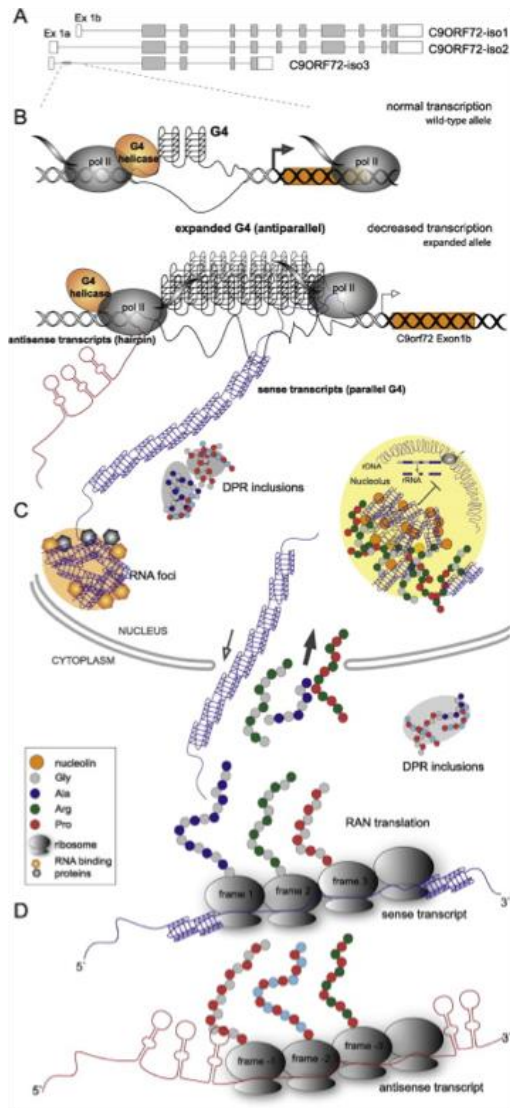


1 (b)

2 Figure 3. Triplet repeat expansion alters mRNA function. (a) In DM1, CTG expansion in the 3'-UTR of
 3 the DMPK gene causes the ensuing mRNA to fold into a large and stable double-stranded hairpin
 4 stabilised by U•U and G•C base pairs, which recruits muscleblind-like (Drosophila) (MBNL1), a
 5 mediator of pre-mRNA alternative splicing regulation. CUG-hairpins also stimulate CUG RNA-binding
 6 protein 1 (CUGBP1) hyperphosphorylation and stabilisation, which alter several events related to
 7 alternative splicing, mRNA translation and mRNA decay. (b) Sequestration of MBNL1 and CUGBP1
 8 activation shift alternative splicing programs from the adult stage towards embryonic-specific
 9 patterns, including activation of exon 5 inclusion of cardiac isoforms of TNNT2 (cTNT) during heart

1 remodeling, exclusion of exon 11 in the insulin receptor (IR) pre-mRNA and inclusion of stop-
 2 containing exon in chloride channel 1 transcripts. Adapted from Lee and Cooper 2009 with kind
 3 permission by Portland Press Ltd. Copyright © the Biochemical Society.

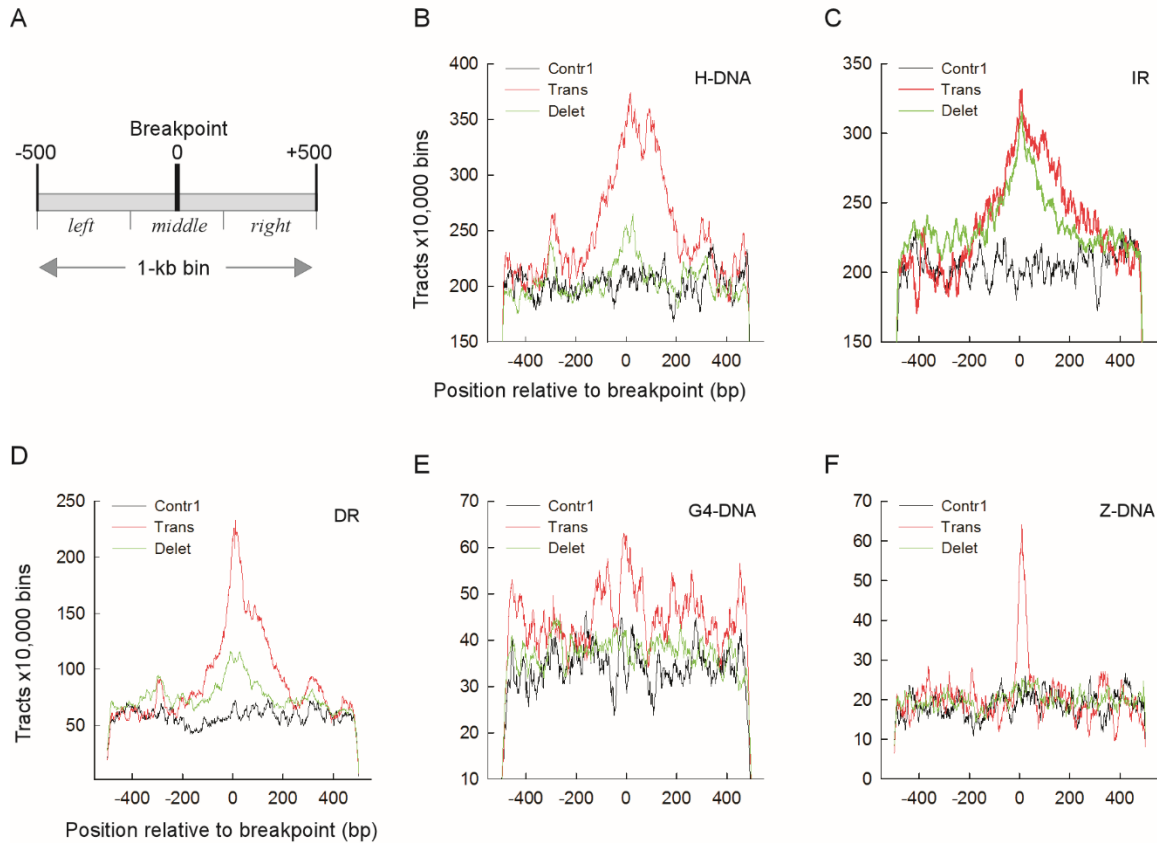
4



5

6 Figure 4. Gain-of-function by expanded G4 DNA-forming repeats at the C9orf72 locus. (a) The
 7 C9orf72 gene is transcribed from two alternative transcription start sites (TSSs; Ex 1b and Ex 1a) in
 8 three gene isoforms. The G4C2 repeats is located in intron 1 (white boxes, UTRs; gray boxes, coding
 9 exons; thin lines, introns) on the non-transcribed strand (thus it is present on the sense RNA)
 10 between Ex 1a and Ex 1b. (b) Normal alleles containing 2-8 G4-forming repeats are transcribed
 11 normally (top). In expanded alleles (bottom) transcription is reduced. The non-transcribed strand
 12 forms an “island” of antiparallel G4 structures. The transcribed strand yields sense RNA with multiple
 13 parallel G4 DNA structures. Antisense transcription also takes place through the island, yielding
 14 antisense RNAs with potential secondary structures. (c) The aberrant transcripts sequester RNA-
 15 binding proteins, forming nuclear protein-RNA foci (left); they also bind nucleolin in nucleoli, where
 16 they prevent biogenesis of new ribosomal RNAs (rRNA; right). (d) Once transported to the
 17 cytoplasm, both sense and antisense transcripts undergo RAN translation in all three possible

1 reading frames, thereby producing dipeptide repeat proteins (DPR) prone to aggregation in the
 2 cytoplasm, nucleus and the nucleolus. Reproduced with permission from Simone et al., 2015.
 3



4
 5 Figure 5. Translocation and deletion breakpoints occur near non-B DNA-forming sequences in cancer
 6 genomes. (a) Schematic of a 1kb interval (bin) with the site of rearrangement (breakpoint) at the
 7 centre and 500 bps of flanking DNA sequence. The genomic location at each breakpoint identified in
 8 cancer patients by high-throughput whole-genome DNA sequencing and resolved at bp resolution
 9 was first mapped to the human reference genome, and 500 bps on either side of each breakpoint
 10 were sought for the occurrence of non-B DNA-forming sequences. (b) Number of triplex DNA-
 11 forming repeats. (c) Number of inverted repeats. (d) Number of direct (tandem) repeats. (e) Number
 12 of G4 DNA-forming repeats. (f) Number of Z-DNA forming repeats. Contr1, 20,222 randomly
 13 generated sites throughout the human genome; Trans, 19,947 chromosomal translocation
 14 breakpoints; Delet, 46,365 deletion breakpoints. In most cases the number of non-B DNA-forming
 15 repeats peaked at the breakpoint position, implying their involvement in triggering DNA strand
 16 breaks that may have elicited the genomic rearrangements. With kind permission from Oxford
 17 University Press (Bacolla et al., 2016).

18