Genome Medicine

MINIREVIEW

Trinucleotide repeats: triggers for genomic disorders?

Piotr Kozlowski†, Krzysztof Sobczak† and Wlodzimierz J Krzyzosiak*

Abstract

Among the various sequence repeats that shape the human genome, trinucleotide repeats have attracted special interest as a result of their involvement in a class of human genetic disorders known as triplet repeat expansion diseases. Recently, long TGG repeat tracts were shown to be implicated in a genomic disorder resulting from chromosome 14q32.2 deletion. Various different mechanisms might trigger this deletion, and looking at the problem from a structural biology perspective may help. Deeper insight into repeated sequences and their features may shed light on the mechanisms involved in this microdeletion and similar genomic rearrangements.

Genomic repeats and human diseases

At least a third of the human genome consists of repetitive sequences of various types, including large segmental duplications, also known as low-copy-number repeats (LCRs), long and short interspersed transposon-derived elements (LINEs and SINEs) and tandem repeats [1]. The tandemly repeated sequences encompass satellites (with repeated units longer than 100 bp), minisatellites (between 100 bp and 10 bp) and microsatellites (with a repeated motif shorter than 10 bp) [2]. The latter, also known as short tandem repeats (STRs) or simple sequence repeats, account for about 3% of the genome. Most of the STR tracts occur in the intergenic regions and introns, but a fraction of them, predominantly trinucleotide repeats (TNRs), also reside in exons and may be beneficial, neutral or deleterious. Among the beneficial roles of TNRs, which contribute about 0.1% to all STR sequences and are often polymorphic in length, is their potential to modulate cellular processes, including transcription splicing and translation [3]. These TNRs include repeats of CGG, CAG and AGG, which are overrepresented in human exons [4]. On the other hand, AAT, AAC and AAG are probably disadvantageous as they are negatively selected in exons [4]. TNR sequences undergo mutations at a very high frequency [5], and this may increase disease risk or trigger disease in specific conditions [6,7].

Over the past two decades our thinking about the links between STRs and human diseases has been dominated by neurological disorders known as trinucleotide repeat expansion diseases (TREDs) [8,9]. There are over 20 diseases that belong to this group, the best known of which are fragile X syndrome (FXS), myotonic dystrophy type 1 (DM1), Huntington's disease (HD) and spinocerebellar ataxias (SCAs). FXS is caused by an expanded CGG repeat located in the 5' untranslated region (UTR) of the fragile X mental retardation 1 gene (*FMR1*); DM1 is triggered by an expanded CUG repeat located in the 3' UTR of the dystrophia myotonica protein kinase gene (*DMPK*); and HD is caused by an abnormally elongated CAG repeat located in the open reading frame of the Huntingtin gene (*HTT*), which is translated to form a polyglutamine tract in the protein (Figure 1a-c). The repeat type and localization determines the mechanism of pathogenesis, which can be impaired transcription (FXS, Figure 1a), transcript toxicity (DM1, Figure 1b) or protein toxicity (HD and SCAs; Figure 1c) [10,11].

Research on the pathogenesis of TREDs includes studies on toxic RNA that triggers alternative splicing alteration in numerous genes linked to the clinical symptoms of DM1 [12,13], and studies on toxic poly-Q proteins that impair many cellular functions [11]. Research on repeat instability mechanisms is also very active, and there are still many challenges ahead [7,8]. The consensus opinion at present is that several processes, including replication, recombination, DNA repair and transcription, contribute to repeat instability and that the formation of unusual non-B-DNA structures formed by the repeats is at the heart of the expansion processes [7,8]. When classified by the size of the underlying mutation, TREDs lie between many genetic diseases resulting from small base substitutions, deletions and

†These authors contributed equally to this work.
*Correspondence: wlodkrzy@ibch.poznan.pl
Laboratory of Cancer Genetics, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland.

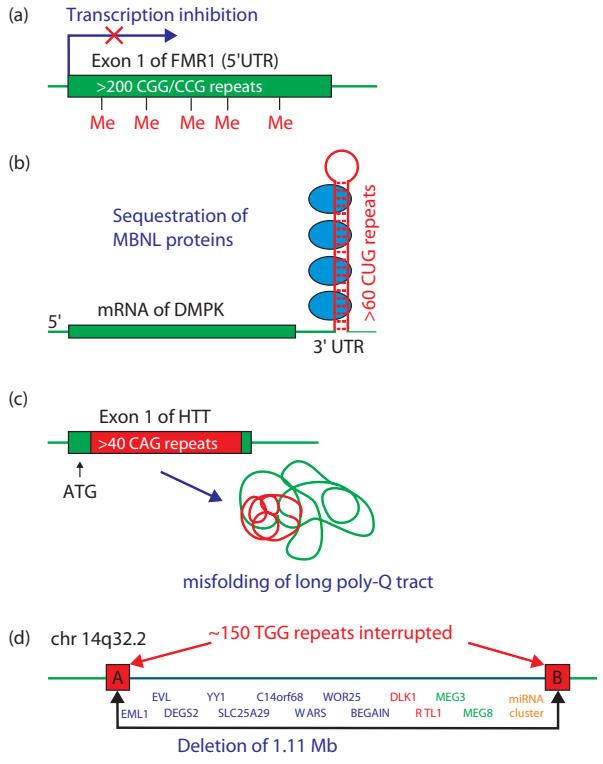

Figure 1. Triplet-repeat-mediated pathological mechanisms of human diseases. (a-c) Diseases associated with the expansion of triplet repeats (TREDs). **(a)** Expansion of CGG/CCG repeats over 200 repeats in exon 1 of the *FMR1* gene located on chromosome X causes methylation of CpG islands in expanded repeats and flanking DNA, which results in the formation of heterochromatin and inhibition of transcription. Loss of *FMR1* expression causes FXS in mutation-carrying males; FXS is thus a recessive disease. **(b)** Expanded CTG repeats (60 to a few thousand) in the 3' UTR of the *DMPK* gene are transcribed but not translated. Long CUG repeat hairpins cause a toxic dominant RNA gain-of-function effect mediated by sequestration of nuclear RNA-binding proteins, such as the alternative splicing regulator *muscleblind-like 1* (MBNL1). There is clear evidence of an RNA gain-of-function effect in at least five TREDs: DM1, DM2 (expanded CCTG repeats), fragile X-associated tremor ataxia syndrome (FXTAS; expanded CGG repeats), Huntington's disease-like 2 (HDL2) and SCA8 (expanded CTG repeats). **(c)** The mutated *HTT* gene with expanded CAG repeats (40 to 100 repeats) in the coding region is transcribed and translated into a toxic protein containing an abnormally long polyglutamine domain. Intracellular aggregation of mutant protein is responsible for the pathogenesis of HD. A similar pathological mechanism is postulated for several dominant disorders known as polyglutamine expansion diseases: seven different spinocerebellar ataxias (SCA1, 2, 3, 6, 7, 8 and 17), dentatorubral-pallidoluysian atrophy (DRPLA) and spinal and bulbar muscular atrophy (SBMA). **(d)** Diseases caused by long TGG repeat tracts. The dominant UPD(14)mat-like phenotype is caused by the deletion of a 1.11 Mb fragment of chromosome 14q32, which is mediated by two interrupted TGG repeat tracts (red boxes A and B). The deleted fragment contains about a dozen protein and short RNA coding genes, including paternally (green) and maternally (red) imprinted genes. The phenotype results from loss of function of two genes, *DLK1* and *RTL1*, and haplo-insufficiency of the others.

insertions and a class of diseases known as genomic disorders, caused by deletions or insertions of tens of thousands to several million base pairs. The group of genomic disorders with identified mutation mechanisms is constantly increasing, with major mechanisms including non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ) and replication fork stalling and template switching (FoSTeS) [14].

TGG repeats trigger recurrent microdeletion

A recently published article [15] shows a link between a TNR sequence and a human genomic disorder related to OMIM #608149. The authors demonstrated that the recurrent 1.11 Mb microdeletion from the long arm of paternal chromosome 14 (14q32.2) is catalyzed by long tracts of interrupted TGG repeats (approximately 500 bp in size) located at both sides of the deletion with 88% sequence similarity (Figure 1d). An identical heterozygous deletion was found in two unrelated patients diagnosed with several clinical phenotypes (such as growth retardation, hypotonia, precocious puberty and mental retardation) characteristic of maternal uniparental disomy (UPD(14)mat). UPD is defined by the inheritance of two copies of a chromosome from only one parent, a mother in this case, and is related to parent-specific imprinting of some genes. The deleted 14q32.2 region harbors 13 protein-coding genes, small nucleolar RNA (snoRNA) and microRNA loci [15] (Figure 1d). Two of these genes, Delta-like homolog 1 (*DLK1*) and retrotransposon-like 1 (*RTL1*), are maternally imprinted (paternally expressed), which explains several disease symptoms [15].

The authors [15] considered several possible deletion mechanisms (Figure 2b). First, the deletion may be mediated by NAHR that occurs between two TGG repeat tracts. Second, it may result from an inherent instability of the repeat and/or from the stable structure that the repeated sequence is very likely to form, and either of these would affect the second and third possible mechanisms, NHEJ and FoSTeS. NAHR is the mechanism that best explains genomic rearrangements in which sites are flanked by highly similar sequences. Most of the recurrent genomic rearrangements that have a common size and fixed breakpoints are thought to occur by NAHR [14]. However, none of the recurrent genomic disorders known so far, perhaps with the exception of some cases of Jacobsen syndrome [16], have recombination hot spots located in triplet repeat tracts. Typically, the NAHR breakpoints are located in LCRs 10 to 300 kb in size that share over 95% similarity [14]. NAHR hotspots are typically 300 to 500 bp in size and contain non-B DNA structures capable of inducing double-stranded DNA (dsDNA) breaks, such as palindromes, DNA transposons and minisatellites but not microsatellites [17]. The STR

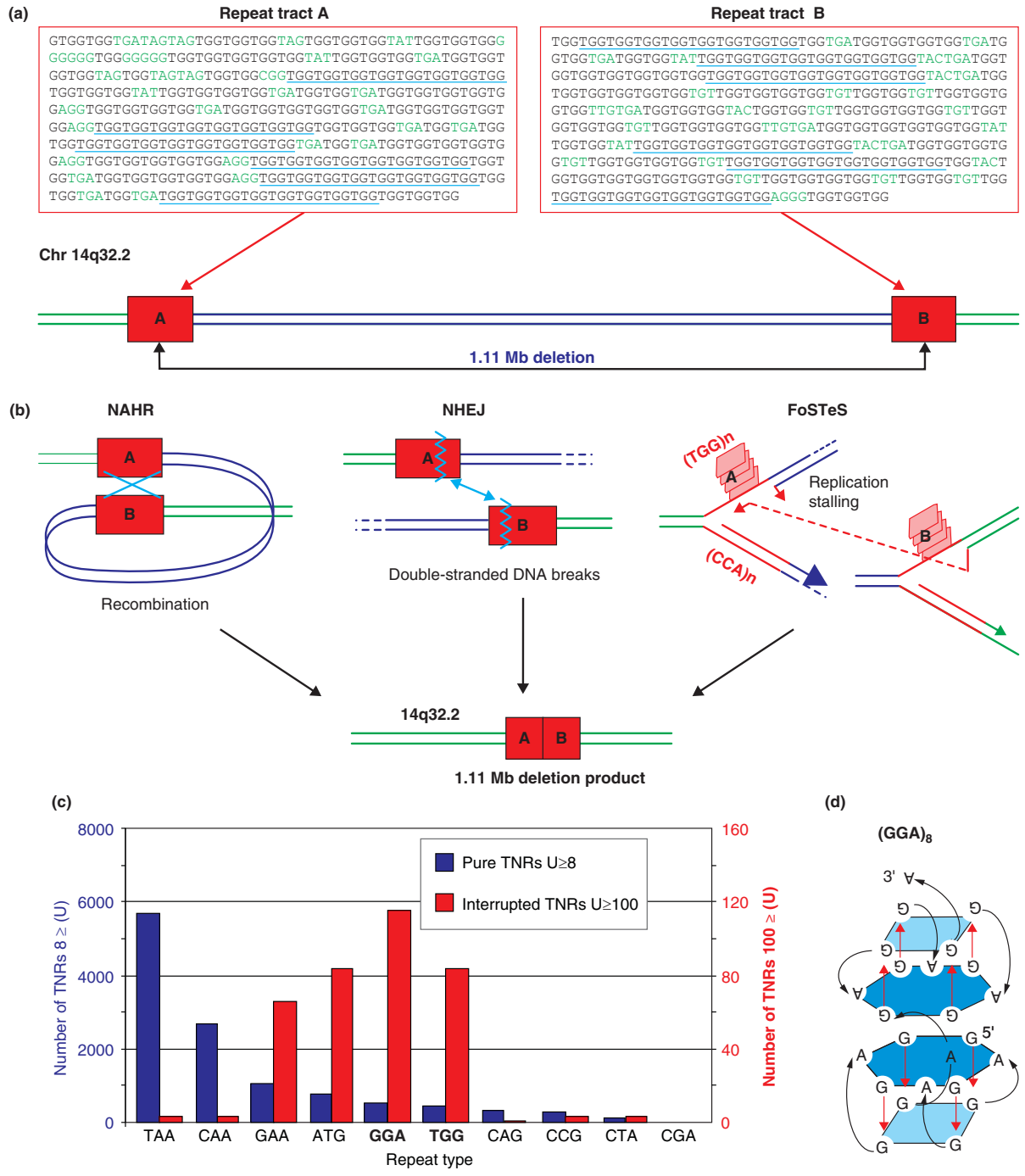

Figure 2. Mechanism of TGG repeat-induced deletion. (a) Schematic representation of the 14q32.2 deletion (blue lines) and proximal and distal breakpoint sequences (red boxes A and B, respectively). Nucleotide sequences of A and B TGG repeat tracts are shown. Green indicates interruptions; pure repeat tracts (of at least 8 repeat units) are underlined. **(b)** Potential mechanisms that can explain the formation of TGG repeat-mediated deletion. NAHR requires homology between breakpoint sequences, NHEJ relies on joining dsDNA breaks induced by DNA structures at breakpoints, and FoSTeS depends on replication stalling and switching of the lagging strand to another replication fork. Both replication stalling and disengagement of the lagging strand can be facilitated by structures formed by template or synthesized DNA strands. **(c)** Frequency of different TNRs in the human genome. Blue bars indicate the number of pure TNR tracts with at least eight repeat units according to [4] (this is the length required for stable G-quadruplex formation); red bars indicate the number of interrupted TNR tracts with at least 100 units (the minimal sequence length required for catalyzing NAHR is 300 bp) according to Simple Repeat track, available on the UCSC Browser (hg18) (our unpublished data). **(d)** G-quadruplex structure formed by eight GGA DNA repeats $(GGA)_8$ [21]. The most 5′ and 3′ nucleotides are shown and arrows indicate direction of DNA strand from 5′ to 3′ end. A similar structure can be expected for TGG repeats based on the results of an RNA study [23].

sequences are typically associated with a second recombination mechanism, NHEJ (Figure 2b), which has evolved to repair dsDNA breaks [17] and as such does not require sequence similarity at breakpoints. A third mechanism, FoSTeS, involves switching of the replicated strand to another replication fork (Figure 2b), which could also happen at TGG repeats [14]. None of these three mechanisms requires TGG repeat expansion, but repeat polymorphisms could modulate deletion frequency.

Structural insight into TGG repeats

A closer inspection of the nucleotide sequences of the TGG repeat segments (Figure 2a) may shed more light on the likelihood of the proposed mechanisms. Both segments (A and B in Figure 2a) contain approximately 60 repeat interruptions (mainly single nucleotide substitutions). The longest uninterrupted TGG repeat is 15 repeat units, and 12 tracts are at least 8 units. Pure repeat tracts of this length probably show only moderate repeat number polymorphism [18]. The repeat interruptions are mostly TGA, TAG and AGG triplets in one repeat tract and TGA, TGT and TAC in the other (Figure 2a). The interrupting triplets may prevent repeated sequences from expansion, which is known to be the case for interrupted CGG and CAG repeats in genes implicated in FXS, SCA1 and SCA2 [19]. Repeat expansions in these genes require the previous loss of repeat interruptions, which are thought to inhibit inter-strand slippage and to suppress intra-strand interaction [7,19]. Bena *et al.* [15] consider the possibility that the TGG repeat tracts are unstable. They demonstrate that TGG repeats are, on average, much longer than any other TNR in the genome. The analysis we have performed using the same constraints (our unpublished work) shows the frequency of TNR tracts in the genome and reveals that AGG and TGG repeats most frequently form the longest tracts of at least 100 units (300 bp), which may facilitate the NAHR mechanism (Figure 2c). Considering only pure repeat tracts of at least 8 units, which may be implicated in repeat instability, the total number of TGG repeats in the genome is similar to that of AGG and much lower than that of TAA and CAA repeats (Figure 2c) [4].

Taking the structural perspective, the repeated sequences within DNA become transiently single-stranded during DNA replication, recombination, repair and transcription, which allows non-B-DNA structure formation and various downstream effects [20]. The repeat interruptions present within the TGG repeats will no doubt influence their ability to form G-quadruplexes and would be likely to diversify the G-quadruplex structures. It is likely that there will be a heterogeneous mixture of structural variants formed by the repeated sequence and their core elements may resemble the G-quadruplex structures described for AGG repeats (Figure 2d) [21]. Notably, the longest repeat tracts of at least 100 units consist of AGG and TGG repeats (Figure 2c), which are capable of forming G-quadruplex structures. For both of these repeat types, the presence of just four repeats is sufficient to form minimal G-quadruplex structures (Figure 2d) that can stack on each other and become more stable. One lesson that can be taken from our analysis of the putative mechanisms underlying the 14q32.2 deletion is that deeper insight into the features of repeated sequences may be needed to identify and better understand the mechanism involved.

The tip of the iceberg or a scarce phenomenon?

Whatever the exact mechanism implicated in the 14q32.2 deletion [15], the involvement of TGG repeat tracts in this deletion cannot be questioned. One important issue that needs to be addressed now is how general this kind of mechanism could be. If NAHR is in operation, similar TNR-mediated genomic rearrangements should be predictable, as was shown earlier for LCR sequences [22]. If stable structure is important, the analysis can be narrowed to repeats having the potential to form G-quadruplex (TGG, AGG and CGG) and hairpin (CNG, GAC and GTC) structures [23,24]. If repeat instability is essential, more attention needs to be paid to the nature, density and localization of the repeat interruptions. Genome-wide copy-number variation discovery studies (for example, [25]) may provide important information on this intriguing question.

Abbreviations

DM1: myotonic dystrophy type 1; FoSTeS: replication fork stalling and template switching; FXS: fragile X syndrome; HD: Huntington's disease; LCR: low-copy-number repeats; NAHR: non-allelic homologous recombination; NHEJ: non-homologous end joining; SCA: spinocerebellar ataxia; STR: short tandem repeat; TNR: trinucleotide repeat; TRED: trinucleotide repeat expansion disease; UPD: uniparental disomy; UTR: untranslated region.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PK contributed genomic analyses, KS contributed structural analyses and WJK wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Ministry of Science and Higher Education, Grant Nos. PBZ-MNiI-2/1/2005, N301-112-32/3910, N302-278937, N302-260938, and Operational Program 'Innovative economy' POIG.01.03.01-00-098/08.

Published: 30 April 2010

References

1. Jasinska A, Krzyzosiak WJ: **Repetitive sequences that shape the human transcriptome.** *FEBS Lett* 2004, **567:**136-141.
2. Voineagu I, Freudenreich CH, Mirkin SM: **Checkpoint responses to unusual structures formed by DNA repeats.** *Mol Carcinog* 2009, **48:**309-318.
3. Kashi Y, King DG: **Simple sequence repeats as advantageous mutators in evolution.** *Trends Genet* 2006, **22:**253-259.
4. Kozlowski P, de Mezer M, Krzyzosiak WJ: **Trinucleotide repeats in human**

genome and exome. *Nucleic Acids Res* 2010, doi:10.1093/nar/gkq127.
5. Hurles M: **How homologous recombination generates a mutable genome.** *Hum Genomics* 2005, **2:**179-186.
6. Hannan AJ: **Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'.** *Trends Genet*, **26:**59-65.
7. Mirkin SM: **Expandable DNA repeats and human disease.** *Nature* 2007, **447:**932-940.
8. Lopez Castel A, Cleary JD, Pearson CE: **Repeat instability as the basis for human diseases and as a potential target for therapy.** *Nat Rev Mol Cell Biol*, **11:**165-170.
9. Orr HT, Zoghbi HY: **Trinucleotide repeat disorders.** *Annu Rev Neurosci* 2007, **30:**575-621.
10. Ranum LP, Day JW: **Pathogenic RNA repeats: an expanding role in genetic disease.** *Trends Genet* 2004, **20:**506-512.
11. Zoghbi HY, Orr HT: **Pathogenic mechanisms of a polyglutamine-mediated neurodegenerative disease, spinocerebellar ataxia type 1.** *J Biol Chem* 2009, **284:**7425-7429.
12. Shin J, Charizanis K, Swanson MS: **Pathogenic RNAs in microsatellite expansion disease.** *Neurosci Lett* 2009, **466:**99-102.
13. Cooper TA, Wan L, Dreyfuss G: **RNA and disease.** *Cell* 2009, **136:**777-793.
14. Zhang F, Gu W, Hurles ME, Lupski JR: **Copy number variation in human health, disease, and evolution.** *Annu Rev Genomics Hum Genet* 2009, **10:**451-481.
15. Bena F, Gimelli S, Migliavacca E, Brun-Druc N, Buiting K, Antonarakis SE, Sharp AJ: **A recurrent 14q32.2 microdeletion mediated by expanded TGG repeats.** *Hum Mol Genet* 2010, doi:10.1093/hmg/ddq075.
16. Jones C, Mullenbach R, Grossfeld P, Auer R, Favier R, Chien K, James M, Tunnacliffe A, Cotter F: **Co-localisation of CCG repeats and chromosome deletion breakpoints in Jacobsen syndrome: evidence for a common mechanism of chromosome breakage.** *Hum Mol Genet* 2000, **9:**1201-1208.
17. Gu W, Zhang F, Lupski JR: **Mechanisms for human genomic rearrangements.** *Pathogenetics* 2008, **1:**4.
18. Rozanska M, Sobczak K, Jasinska A, Napierala M, Kaczynska D, Czerny A, Koziel M, Kozlowski P, Olejniczak M, Krzyzosiak WJ: **CAG and CTG repeat polymorphism in exons of human genes shows distinct features at the expandable loci.** *Hum Mutat* 2007, **28:**451-458.
19. Pearson CE, Eichler EE, Lorenzetti D, Kramer SF, Zoghbi HY, Nelson DL, Sinden RR: **Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation.** *Biochemistry* 1998, **37:**2701-2708.
20. Lin Y, Dent SY, Wilson JH, Wells RD, Napierala M: **R loops stimulate genetic instability of CTG.CAG repeats.** *Proc Natl Acad Sci USA* 2010, **107:**692-697.
21. Matsugami A, Okuizumi T, Uesugi S, Katahira M: **Intramolecular higher order packing of parallel quadruplexes comprising a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad of GGA triplet repeat DNA.** *J Biol Chem* 2003, **278:**28147-28153.
22. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE: **Segmental duplications and copy-number variation in the human genome.** *Am J Hum Genet* 2005, **77:**78-88.
23. Sobczak K, Michlewski G, de Mezer M, Kierzek E, Krol J, Olejniczak M, Kierzek R, Krzyzosiak WJ: **Structural diversity of triplet repeat RNAs.** *J Biol Chem* 2010, doi:10.1074/jbc.M109.078790.
24. Bacolla A, Larson JE, Collins JR, Li J, Milosavljevic A, Stenson PD, Cooper DN, Wells RD: **Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties.** *Genome Res* 2008, **18:**1545-1553.
25. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J; The Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME: **Origins and functional impact of copy number variation in the human genome.** *Nature* 2009, **464:**704-712.

doi:10.1186/gm150
Cite this article as: Kozlowski P, *et al*.: Trinucleotide repeats: triggers for genomic disorders? *Genome Medicine* 2010, **2:**29.