The G-rich Repeats in *FMR1* and *C9orf72* Loci Are Hotspots for Local Unpairing of DNA

Manar Abu Diab,*,†,1 Hagar Mor-Shaked,*,†,1 Eliora Cohen,*,† Yaara Cohen-Hadad,*,† Oren Ram,‡ Silvina Epsztejn-Litman,* and Rachel Eiges*,†,1,2

*Stem Cell Research Laboratory, Medical Genetics Institute, Shaare Zedek Medical Center, Jerusalem 91031, Israel, †The Hebrew University School of Medicine, Jerusalem 91120, Israel, and ‡Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

ORCID ID: 0000-0001-8139-0933 (R.E.)

ABSTRACT Pathological mutations involving noncoding microsatellite repeats are typically located near promoters in CpG islands and are coupled with extensive repeat instability when sufficiently long. What causes these regions to be prone to repeat instability is not fully understood. There is a general consensus that instability results from the induction of unusual structures in the DNA by the repeats as a consequence of mispairing between complementary strands. In addition, there is some evidence that repeat instability is mediated by RNA transcription through the formation of three-stranded nucleic structures composed of persistent DNA:RNA hybrids, concomitant with single-strand DNA displacements (R-loops). Using human embryonic stem cells with wild-type and repeat expanded alleles in the *FMR1* (CGGs) and *C9orf72* (GGGGCCs) genes, we show that these loci constitute preferential sites (hotspots) for DNA unpairing. When R-loops are formed, DNA unpairing is more extensive, and is coupled with the interruptions of double-strand structures by the nontranscribing (G-rich) DNA strand. These interruptions are likely to reflect unusual structures in the DNA that drive repeat instability when the G-rich repeats considerably expand. Further, we demonstrate that when the CGGs in *FMR1* are hypermethylated and transcriptionally inactive, local DNA unpairing is abolished. Our study thus takes one more step toward the identification of dynamic, unconventional DNA structures across the G-rich repeats at *FMR1* and *C9orf72* disease-associated loci.

KEYWORDS Unstable repeat expansions; *FMR1*; *C9orf72*; R-loops; single-strand DNA displacements

PATHOLOGICAL mutations involving DNA microsatellite repeat expansions are responsible for over 40 different neurodevelopmental, neurodegenerative, and neuromuscular diseases. All result from expanded tracts of repeated DNA sequences that become unstable beyond a critical length [for comprehensive reviews see Gatchel and Zoghbi (2005), Pearson *et al.* (2005)]. In this group of conditions, one class arises from particularly large repeat expansions (hundreds to thousands of repeat copies) such as fragile X syndrome (FXS) and C9-related amyotrophic lateral sclerosis and/or frontotemporal dementia (C9/ALS-FTD). FXS is one of the most common heritable forms of cognitive impairment and is caused by a CGG repeat expansion (>200 repeats) in the X-linked *FMR1* gene (Oberle *et al.* 1991; Verkerk *et al.* 1991), whereas C9/ALS-FTD is the most commonly known cause of amyotrophic lateral sclerosis or frontotemporal dementia and results from a GGGGCC repeat expansion (>30 repeats) in the *C9orf72* gene (DeJesus-Hernandez *et al.* 2011; Dols-Icardo *et al.* 2014). Large repeat expansions, like those in FXS and C9/ALS-FTD, are located in noncoding regions of genes (in *FMR1* and *C9orf72* at the 5′-UTR of the gene). They are typically positioned next to CpG island promoters, and present extensive repeat instability when sufficiently long. Although the timing, pattern, and tissue selectivity of somatic repeat instability varies across repeat-associated pathologies, for each condition, aberrant mispairing between complementary strands as exhibited by single-strand DNA (ssDNA) displacements is assumed to provide the initial trigger for instability. There is a general consensus

doi: https://doi.org/10.1534/genetics.118.301672
Manuscript received August 25, 2018; accepted for publication October 15, 2018; published Early Online November 5, 2018.
Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7295477.
[1]These authors contributed equally to this work.
[2]Corresponding author: Stem Cell Research Laboratory, Medical Genetics Institute, Shaare Zedek Medical Center, Affiliated with The Hebrew University School of Medicine, PO Box 3235, 12 Bayit St., Jerusalem 91031, Israel. E-mail: rachela@szmc.org.il

that ssDNA displacements promote instability through aberrant DNA repair, although local disruption in DNA replication has also been implicated [for comprehensive reviews see Pearson *et al.* (2005), Mirkin (2007)]. Regardless of the mechanism, all repeat instability models are based on the formation of hard-to-process noncanonical structures generated by the unpaired DNA (Cleary *et al.* 2002; Panigrahi *et al.* 2005, 2012; Salinas-Rios *et al.* 2011; Axford *et al.* 2013; Slean *et al.* 2013). These structures are then resolved by the addition (expansions) or deletion (contractions) of repeats (Usdin and Woodford 1995; Samadashwily *et al.* 1997; Pearson *et al.* 1998a; Cleary *et al.* 2002; Nichol Edamura *et al.* 2005), eventually leading to repeat size mosaicism in the patients' somatic cells. Furthermore, there is experimental evidence to support the involvement of R-loops in mediating repeat instability (Panigrahi *et al.* 2005; Grabczyk *et al.* 2007; Lin *et al.* 2010; Salinas-Rios *et al.* 2011; Reddy *et al.* 2014; Slean *et al.* 2016; Su and Freudenreich 2017). R-loops are three-stranded nucleic acid structures that are composed of persistent DNA:RNA hybrids (Salinas-Rios *et al.* 2011). They are formed naturally as the result of reannealing of nascent RNA transcripts with the template DNA as soon as they exit the transcription bubble, concomitantly with the formation of ssDNA displacements [for a comprehensive review see Jonkers and Lis (2015)]. They are typically formed by sequences with a strong positive G/C skew (G-clusters in the nontemplate DNA strand) next to transcriptionally active promoters (Ginno *et al.* 2012), and play a central role as key intermediates in a range of fundamental cellular processes. However, R-loops can be a threat to the cell since they can lead to genome instability [reviewed by Aguilera and Garcia-Muse (2012)]. One model implicating R-loops in the enhancement of repeat instability argued that they act by promoting complex noncanonical structures, such as hairpins and G-quadruplexes (G4) by the unpaired DNA in the R-loop (Gray *et al.* 2014). Although studies have provided evidence for the formation of R-loops at the *FMR1* and *C9orf72* loci (Colak *et al.* 2014; Groh *et al.* 2014; Loomis *et al.* 2014; Kumari and Usdin 2016; Esanov *et al.* 2017), the existence of hairpins/G4 structures at those regions *in vivo* remains undocumented. Here, we finely characterize and precisely map R-loops and ssDNA displacements across and near the repeats at the *FMR1* and *C9orf72* loci *in vivo* to better understand the propensity of these loci to become highly unstable. The *FMR1* and *C9orf72* repeats are particularly pertinent to this type of study since they constitute preferred sites for R-loop initiation and are predicted by *in vitro* studies to form complex secondary structures when unpaired.

Using human embryonic stem cells (hESCs) with wild-type and expanded alleles in the FXS or C9/ALS genes (which most resemble early human embryonic cells and are often transcriptionally active; Eiges *et al.* 2007; Avitzour *et al.* 2014; Cohen-Hadad *et al.* 2016), we provide *in vivo* evidence that *FMR1* and *C9orf72* repeats are hotspots for local unpairing of DNA regardless of repeat tract length. Furthermore, we show that in *FMR1* when R-loops are formed, DNA unpairing is more extensive, and is coupled with the induction of double-strand DNA (dsDNA) interruptions on the nontranscribing (G-rich) DNA strand. These structures are predicted to drive instability when the repeats expand. Finally, we provide evidence that when the CGGs in *FMR1* are hyper-methylated and transcriptionally inactive, local DNA unpairing is abolished. This study thus takes an additional step toward the identification of dynamic, unconventional DNA structures across the G/C-rich repeats at the *FMR1* and *C9orf72* genes *in vivo*. Their potential involvement in promoting repeat instability is discussed.

Materials and Methods

Southern blot analysis

Genomic DNAs (10–25 μg) were digested with *Eco*RI and the methylation sensitive restriction enzyme *Eag*I [New England Biolabs, Beverly, MA (NEB)], separated on 0.8% agarose gels, blotted onto Hybond N+ membranes (Amersham, Piscataway, NJ), and hybridized with a PCR Dig-labeled probe (primers: 5′-GCT AGC AGG GCT GAA GAG AA-3′ and 5′-CAG TGG AGC TCT CCG AAG TC-3′).

Real-time quantitative RT-PCR/PCR

Total RNA was isolated from the cells by TRI Reagent extraction, then 1 μg RNA was reverse-transcribed by random hexamer priming and MultiScribe reverse transcriptase (Applied Biosystems, ABI). For pre-messenger RNA expression levels, amplification was performed according to Avitzour *et al.* 2014.

Real-time quantitative RT-PCR by TaqMan

Custom made TaqMan-based expression assays (ABI) for *FMR1* (HS00924540) and *GUS* (HS99999908) were used according to manufacturer's protocol. *GUS* was used as housekeeping gene control for normalization of ΔΔCt mean values.

DNA:RNA immunoprecipitation assay

DNA:RNA immunoprecipitation (DRIP) assay was performed as previously described (Boque-Sastre *et al.* 2017). Briefly, total nucleic acids were extracted by SDS/Proteinase K treatment at 37°, followed by phenol-chloroform extraction and ethanol precipitation. DNA was fragmented using *Hin*dIII, *Eco*RI, *Bsr*GI, *Xba*I, and *Ssp*I, and pretreated, or not, with recombinant RNase H (#B0297S; NEB) overnight. Then, 4 μg of digested DNA was immunoprecipitated with 10 μg of S9.6 antibody (kindly provided by Clinton E. Leysath, National Institutes of Health, or commercially available from Kerafast) overnight. The pulled-down material (with and without RNase H treatment) and 1% input DNA were then subjected to quantitative PCR analysis. *EGR1* and *RPL13A* were used as negative and positive controls, respectively. Fold enrichments are calculated by the $2^{-\Delta CT}$ formula after adjusting input Ct values. Adjusted Ct_{INPUT} was calculated by

Ct_{INPUT} − 6.644 (where 6.644 represents the correction for 1:100 dilution of input sample relative to bound material). ΔCt = Ct_{BOUND} − Adjusted Ct_{INPUT}. For each cell type (+/− RNAse H1), mean values were determined based on at least three independent DRIP experiments. Quantitative PCR assay was performed in triplicate. All designed primers were calibrated to ensure optimal conditions for amplification. Real-time PCR was performed using Power SYBR Green Master Mix (ABI), on an ABI 7900HT instrument (primers are listed in Table S2).

Colony bisulfite footprinting analysis

Bisulfite footprinting was carried out using a previously reported method (Yu *et al.* 2003). Native genomic DNA (1 μg) was modified by bisulfite treatment (EZ DNA methylation Kit; Zymo Research), with a slight modification. DNA was treated with sodium bisulfite under nondenaturing conditions (37° overnight, in dark conditions). Putative R-loop regions were PCR-amplified with FastStart DNA polymerase (Roche), using a pair of converted and unconverted primers (for R-loops boundary detection) or two unconverted primers (for ssDNA displacement detection). All primers and primer melting temperatures (Tm) are listed in Table S2. For ssDNA displacement at *FMR1*, GC-RICH solution was added to the PCR mix, according to the manufacturer's instructions (Roche). Amplified products were cloned, and single colonies were analyzed for cytosine conversions by direct sequencing (ABI 3130), using the BigDye Terminator v3.1 Cycle Sequencing Kit.

Bisulfite footprinting by deep-sequencing

Library construction and run: First, PCR products were purified using SPRI beads (X1), and then a second PCR was performed to add Illumina adapters with the indices. PCR products were cleaned again using SPRI beads (X1), and eluted in a 25 μl elution buffer. To confirm the length and purity of the library, 1 μl was loaded onto an Agilent TapeStation 2200 using the D1000 ScreenTape assay (Agilent Technologies, Santa Clara, CA). Library concentration was measured using the Qubit dsDNA High Sensitivity Assay Kit. The libraries were normalized to 4 nM and loaded on a MiSeq (Illumina, San Diego, CA) using the MiSeq Reagent Kit v2 (500 cycles) with the following conditions: 250 × 2 in a final concentration of 8 pM and 20% phiX spike-in.

Data analysis: To align each read to a reference sequence, we wrote a simple Perl script, which first filtered out PCR errors with reads containing inaccurate repeat numbers. The reference sequences were manually prepared according to the known repeat tract length for each sample. Next, we separated templates from nontemplate reads (based upon the conversion pattern at the repeats) to generate two separate matrices containing coded information on each nucleotide according to the reference sequence. In addition, a third matrix was generated, containing the remaining reads. The coded information is as follows: non-C nucleotide marked as 0; unconverted C (dsDNA structure) marked as 1; converted C (ssDNA structure) marked as 2 if the conversion was to T (correct conversion), and marked as 3 if the conversion was to A or G, most probably due to sequencing errors, which were found to be uncommon. Reads with over 10 sequencing errors (marked as 3) were excluded. For clustering and heatmaps, we randomly selected 10,000 reads using the R statistical gplots library, and generated heatmaps using the heatmap.2 function. At *FMR1*, the total number of compatible reads was <10,000, and therefore the heatmaps were generated from the entire population of molecules.

Data and software availability

Supplemental File S1 contains Southern blot analysis of unmethylated FX hESC lines, (Figure S1A) and relative expression levels of FMR1 RNA in the hESCs by TaqMan quantitative RT-PCR (Figure S1B). File S2 contains bisulfite DNA sequencing data at the boundaries of the ssDNA displacements in XY hESC with premutation (55 < CGGs < 200; PM-ES-2) and unmethylated full mutation (CGGs > 200 repeats; uFM-ES-3) alleles. File S3 contains methylation-sensitive quantitative assay of skewed X-inactivation test in uFM-ES-2 hESCs. File S4 contains bisulfite footprinting analysis by deep-sequencing across the C9orf72 repeats in wild-type hESCs. File S5 contains bisulfite footprinting analysis around the transcription start site (TSS) in other loci. Table S1 contains all the required data regarding the different hESC lines used. Table S2 contains the sequences of all primers used for DRIP, colony bisulfite footprinting, next-generation sequencing bisulfite footprinting, and SNP analyses.

Supplemental files are available at National Center for Biotechnology Information, Gene Expression Omnibus (NCBI GEO) and GitHub. The raw (fastq files) and analyzed (text files) data for the bisulfite footprinting analysis by deep-sequencing in *FMR1*, *C9orf72*, and *RPL13A*, were deposited at the NCBI GEO under accession number GSE111743. Code used to generate the simulated data can be found at GitHub (https://github.com/RachelEiges/Bisulfite_Footprinting). The derivation and use of mutant hESCs was performed in compliance with protocols approved by the Ethics Committee of Shaare Zedek Medical Center (Institutional Review Board approval no. 87/07).

FASTQ data are provided for wild-type alleles in *FMR1*, FMR1_WT-ES-4 for unmethylated (WT-ES-4), and FMR1_unFMES-2 for methylated (uFM-ES-2) alleles; C9orf72_WT-ES-1, C9orf72_WT-ES-5, *C9orf72*_C9-ES1, C9orf72_C9-ES2 for wild-type alleles in C9orf72; four different cell lines representing five different alleles (WT-ES-1, heterozygote with two or five repeats; WT-ES-5, homozygote with two repeats; C9-ES-1, five repeats only; and C9-ES-2, two repeats only); and *RPL13A*_WT-ES-4 for nonrepetitive R-loop forming gene RPL13A (WT-ES-4). The processed data for each FASTQ file are provided as TXT documents: FMR1_WT-ES-4_antisense_conv_tab, FMR1_WT-ES-4 _sense_ conv_tab, FMR1_un-FMES-2_antisense_con_tab, FMR1_

un-FMES-2_sense_con_tab, C9orf72_WT-ES-1_antisense_con_tab, C9orf72_WT-ES-5_2rep_antisense_con_tab, C9orf72_WT-ES-5_ 5rep_antisense_con_tab, C9orf72_C9-ES1_antisense_con_tab, C9orf72_C9-ES2_antisense_con_tab, and RPL13A_WT-ES-4_TXT. Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7295477.

Results

R-loops are regularly formed at the 5′-UTRs of FMR1 and C9orf72 in hESCs

To validate the presence of R-loops at the 5′-UTR of *FMR1* in hESCs where the repeats are located, we conducted DRIP experiments using a monoclonal antibody (S9.6) that recognizes DNA:RNA hybrids. For this purpose, we used hESCs harboring wild-type or/and mutant *FMR1* repeat expansions (Eiges *et al.* 2007; Avitzour *et al.* 2014). Similar to adult fibroblasts from rare subjects with an unmethylated full expansion, these cells are frequently *FMR1* gene-active and often present extensive instability [mainly, but not limited to contractions; see Figure S1A and Eiges *et al.* (2007), Avitzour *et al.* (2014)]. We show that R-loops are regularly formed at the *FMR1* locus. Pretreatment of the samples with RNase H verified the specificity of the pulldown assay, indicating significant enrichments for R-loops in the wild-type and expanded alleles, although to a lesser extent in wild-type alleles (Figure 1A). This cannot be attributed to a difference in transcription levels between the alleles, given the upregulation of *FMR1* transcripts in the premutation (Tassone *et al.* 2007), but not the full mutation alleles (Figure S1B). These results on the accumulation of R-loops in the 5′-end of the *FMR1* gene are consistent with reports on somatic cells with an unmethylated full mutation (Kumari and Usdin 2016) and FXS patient cells following gene reactivation with 5-azadC (Groh *et al.* 2014).

To explore whether R-loop formation across the CGGs in *FMR1* represents a more general feature of unstable microsatellite expansions within CpG islands, we extended our study to the *C9orf72* locus (autosomal dominant inheritance). *C9orf72* harbors a GGGGCC repeat expansion in the first intron of the gene (termed the C9 mutation) between noncoding exons 1a and 1b (DeJesus-Hernandez *et al.* 2011; Dols-Icardo *et al.* 2014). To explore whether R-loops are regularly formed at that region and roughly estimate their levels, we used DRIP analysis in the wild-type and C9-affected hESCs (Figure 1B, (Cohen-Hadad *et al.* 2016)). Albeit low, R-loop enrichments were found at equal levels in wild-type and C9 mutant cells (Figure 1B), providing evidence for their existence in the *C9orf72* endogenous locus as well.

Fine mapping of ssDNA in the R-loop

A major drawback of DRIP assay is its limited resolution, which is caused by the use of restriction enzyme digestion. Therefore, we applied a complementary system that was developed to finely map the exact location of specific R-loop–forming regions in a given locus (Yu *et al.* 2003). Basically, this approach involves treatment of native genomic DNA with bisulfite reagent under nondenaturing conditions. The sodium bisulfite reaction specifically converts cytosine (C) to uracil [which is later replaced by thymine (T)] in ssDNA alone. This makes for easier identification of unpaired DNA displacements by single colony or deep-sequencing (for the principles of the bisulfite footprinting approach, see Figure 2A).

We used R-loop prediction software to narrow our search to a ~1 kb putative region along the 5′-UTR of *FMR1* (Jenjaroenpun *et al.* 2015, 2017). We then searched experimentally for three-way junctions upstream and downstream to the repeats, where the nontranscribing CGG strand is expected to change from a double-strand to a single-strand conformation, and vice versa. An enrichment method with one converted primer (matching bisulfite-modified DNA) and one conventional primer (matching unmodified DNA) made it possible to preferentially amplify DNA molecules that spanned across the three-way junctions representing the tips of the DNA:RNA hybrids. By focusing on the ends rather than on the entire length of the DNA displacements, we were able to bypass the technical difficulty of amplifying the expanded CGG repetitive sequence. This enabled us to show indirectly that ssDNA displacements consistently initiate at or near the TSS of *FMR1* (60–100 bp before the repeats) (Figure 2B and Figure S2). This is consistent with the known overlap between the TSS and 5′ boundary of promoter-associated R-loops (Chen *et al.* 2017; Dumelie and Jaffrey 2017), and is in line with an earlier study by Loomis *et al.* (2013) in wild-type and premutation (PM, intermediate size expansion, $55 < \text{CGGs} < 200$) alleles. Next, we searched for termination sites downstream to the repeats, and identified a single termination site in intron 1, ~400 bp beyond the CGGs in all cell types examined, including the XY wild type, premutation, and unmethylated full mutation (uFM, $\text{CGGs} > 200$) alleles (Figure 2B and Figure S2). The bisulfite footprinting analysis in *FMR1* therefore clearly illustrates that these DNA displacements systematically occur at specific positions in all cell types, and hence are common to both normal and expanded alleles, thus providing preferable sites for R-loop formation. Together with the DRIP data, the strong G/C skew and high G-clusters that exist immediately downstream to the TSS where the repeats reside, strongly suggest that these junctions constitute the edges of an R-loop. If so, our findings imply that the DNA:RNA hybrids spread much further downstream beyond the CGG repeats, and end in intronic sequences. In other words, they indicate that the R-loops in *FMR1* are cotranscriptionally formed between nascent RNA and DNA.

To further corroborate these findings, we searched for the 5′ and 3′-ends of the ssDNA displacements that cooccur with the hybrids at the *C9orf72* locus using R-loop prediction software followed by bisulfite footprinting analysis. Applying the same approach as for *FMR1*, preferential amplification of

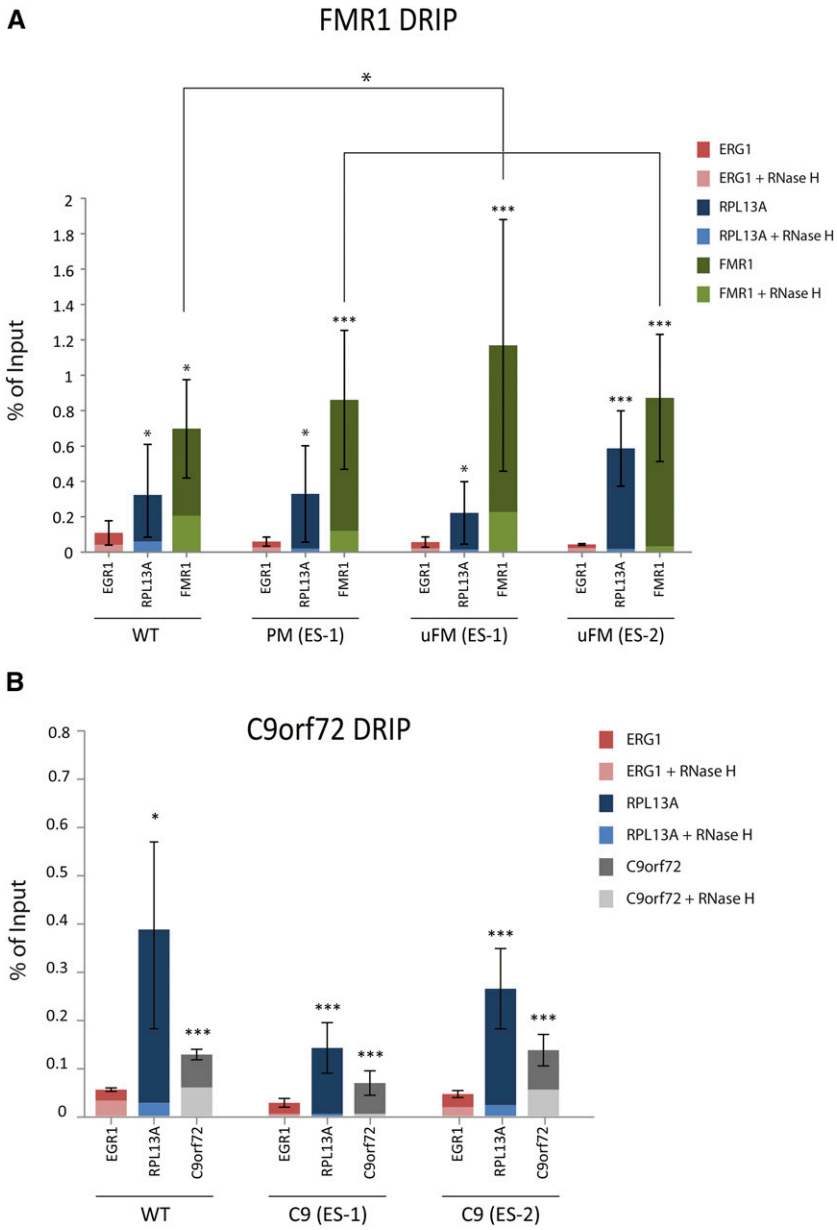


Figure 1 R-loop enrichment at the *FMR1* and *C9orf72* genes by DRIP analysis. (A) R-loop enrichments were determined by DRIP-quantitative PCR with the S9.6 antibody in the *FMR1* locus (green bars), *RPL13A* (positive gene control, blue bars), and *ERG1* (negative control gene, red bars), in hESC with wild-type (WT-ES-1, WT-ES-2), premutation (PM, 55 < CGGs < 200, PM-ES-1), and unmethylated full mutation alleles (uFM, CGGs > 200, uFM-ES-1, uFM-ES-2). The uFM-ES-2 is an XX cell line with skewed X-inactivation so that the wild-type allele is always inactive (Avitzour *et al.* 2014). DRIP samples were pretreated (light green, light blue, and light red, respectively) or untreated (dark green, dark blue, and dark red, respectively) with RNase H. The addition of RNase H was used as a control for the pulldown assay. DRIP values are presented as percentage of input (paired *t*-test, * $P <$ 0.05, *** $P < 0.001$). Bars represent mean values of ±SD (n = 3–5). Note the significant difference in enrichment levels between wild-type and CGG-expanded alleles (PM and uFM). (B) R-loop enrichments were determined at the C9 locus (gray bars), *RPL13A* (positive control, blue bars), and *ERG1* (negative control gene, red bars), as described in A, in control (WT-ES-1) and C9 mutant hESCs (C9-ES-1, C9-ES-2) (Cohen-Hadad *et al.* 2016). DRIP values are presented as percentage of input (paired *t*-test, * $P < 0.05$, *** $P < 0.001$). Bars represent mean values of ±SD (n = 3–5). For cell line specifications and expansion size see Table S1. WT, wild type.

hybrid junctions successfully identified ssDNA displacements initiating nearly 70 bp before the GGGGCC repeats and terminating ~700 bp further downstream into intron 1 (Figure 3). Exploiting an informative SNP downstream to the GGGGCCs (NCBI refSNP database, rs2282240), which permits differentiation between alleles in C9 hESCs, provided further evidence that the DNA displacements are equally formed by the normal and expanded alleles within affected cell lines (Figure 3). We assume that the extrusion sites represent the tips of one long R-loop. A comparison of the initiation and termination sites of the DNA displacements between affected *vs.* nonaffected cells and between normal *vs.* expanded alleles within affected cell lines provided further evidence that the size of the hybrids is dictated by the number of repeats. Thus, along with the *FMR1* data, we suggest that persistent DNA:nascent RNA hybrids are a general feature of G-rich repetitive sequences located proximal to CpG island promoters.

DNA unpairing at/near the FMR1 repeats

To firmly establish that the boundaries of the ssDNA displacements represent one long continuous molecule, we mapped the 5′-end of the hybrids using a different reverse primer that anneals to the bisulfite-converted sequence downstream to the repeats. We analyzed a selection of ssDNA molecules that coincide with the formation of R-loops and span the repeats. The analysis was limited to the *FMR1* locus in wild-type hESCs (WT-ES-1) because the CGGs are extremely hard to PCR amplify, particularly when they are expanded. We found that the ssDNA displacements were consistently interrupted

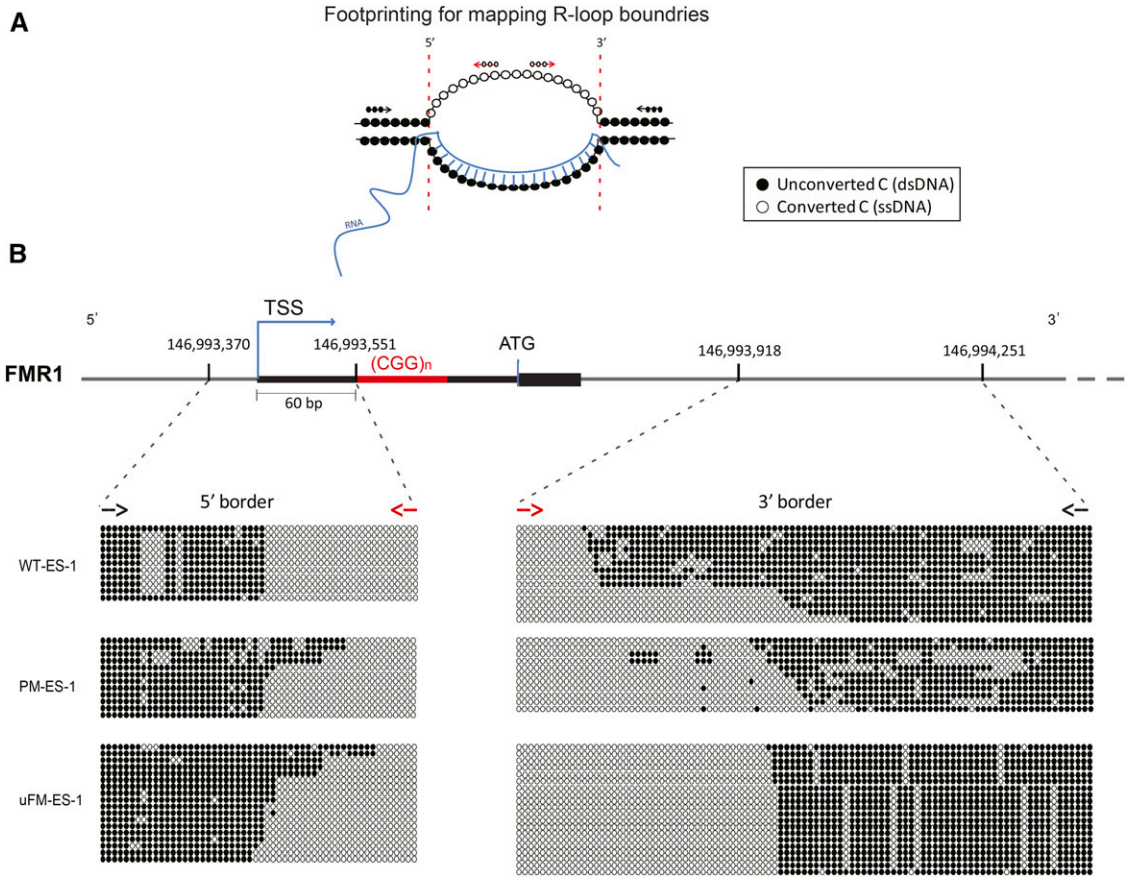


Figure 2 Mapping single-strand DNA displacements on the nontemplate (G-rich) DNA of *FMR1* by bisulfite DNA footprinting. (A) Graphic demonstrating the principles of bisulfite footprinting analysis for mapping the boundaries of single-strand DNA displacements under nondenatured conditions. Each circle represents a cytosine site. Black circles represent bisulfite unconverted C (indicating double-strand DNA conformation) and white circles represent bisulfite converted C to T (indicating single-strand DNA conformation). RNA is indicated by a blue line. Black and red arrows correspond to unconverted and converted primers, respectively. Red dashed lines represent the 5′ and 3′ boundaries. (B) Schematic representing the 5′-UTR of *FMR1* including the TSS (blue) and the CGG repeats (red). Black and red dashed arrows designate unconverted and converted primers, respectively. Bisulfite DNA sequencing data at the boundaries of the single-strand DNA displacements in XY hESC with wild-type (WT-ES-1), premutation (PM-ES-1), and unmethylated full mutation (uFM-ES-1) alleles are presented. Each row represents a single DNA molecule and each circle represents a single cytosine site. Black circles represent bisulfite unconverted C (indicating double-strand DNA conformation) and white circles represent bisulfite converted C to T (indicating single-strand DNA conformation).

at the first half of the repeats by a short dsDNA segment (depicted by the black squares within the large blocks of white squares in Figure 4, A and B). Although we could not entirely rule out the possibility of the formation of multiple R-loops along the region, we speculate that these short double-strand interruptions represent the propensity to generate hairpins, G4s, or any other double-strand secondary structure that interferes with the conversion of C to T by bisulfite treatment.

We next footprinted the nontemplate DNA molecules that are uncoupled with R-loops in two unaffected cell lines (WT-ES-1 and WT-ES-4), using unconverted primers. This was done to verify that in the absence of R-loops, the DNA strands are completely paired along the entire region. Unexpectedly, we found that the double-strand conformation was repeatedly interrupted by short ssDNA segments at the CGG repeats (depicted by the white squares within the large blocks of black squares in Figure 4C). This also permitted the identification of molecules from the C-rich template DNA strand, thus verifying that the template DNA is paired with nascent RNA along the entire region. Here again, we found that the double-strand conformation was repeatedly interrupted by ssDNA segments at the CCG repeats (depicted by white squares within the large blocks of black squares in Figure 4D). These results are indicative of a comprehensive feature of restricted DNA unpairing at the repeats.

To test for correlations between the AGG-repeat interruptions and the formation and position of the DNA displacements, we compared two different wild-type hESC lines, since one has a pair of AGG interruptions (WT-ES-1) whereas the other does not (WT-ES-4). In the cells that carry AGG interruptions, the AGGs were located in both single- and double-strand positions (Figure 4). By contrast, in the cells that lack interruptions, the DNA displacements were longer and more homogeneous (continuously white squares) (Figure 4, C and D). These differences may possibly hint at the functional

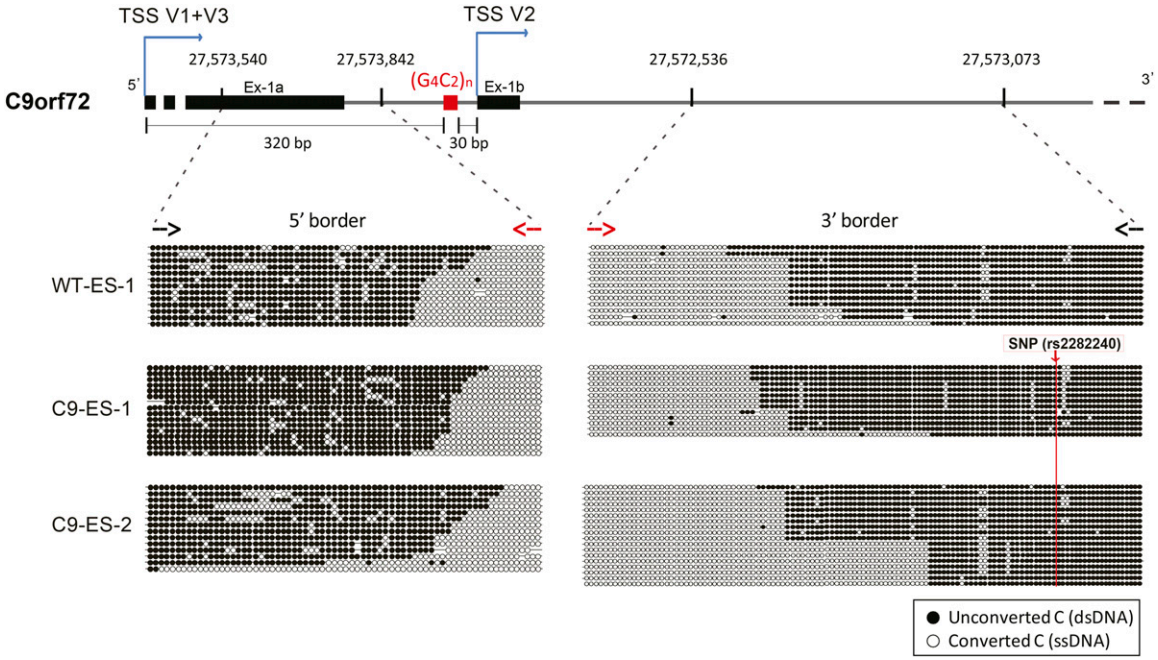


Figure 3 Single-strand DNA displacements on the nontemplate (G-rich) DNA of *C9orf72* by bisulfite DNA footprinting. Schematic representing the 5′-UTR of *C9orf72* including two alternative TSS sites (blue) and the GGGGCC repeats (red). Bisulfite DNA sequencing data at the boundaries of the single-strand DNA displacements in wild-type hESCs (WT-ES-1), and C9 mutant hESCs (C9-ES-1, C9-ES-2) are presented. Each row represents a single DNA molecule and each circle represents a single cytosine site. Black circles represent bisulfite unconverted C (indicating double-strand DNA conformation) and white circles represent bisulfite converted C to T (indicating single-strand DNA conformation). The location of an informative SNP (NCBI refSNP database, rs2282240), which was utilized for validating the equal contribution of the wild-type and expanded alleles in C9-ES-1 and C9-ES-2, is marked in red.

importance of the AGG interruptions in restricting secondary structures at the repeats.

Colony bisulfite sequencing may misrepresent the entire population of molecules. This is due, in part, to PCR duplicates and the low number of sequenced molecules. Therefore, to provide a better representation of the DNA conformations that may form in this region, we extended the analysis to bisulfite deep-sequencing using unconverted primers that flank the repeats. Sequencing reads were processed using a custom analysis script. Reads were excluded from the analysis if they were shorter than 250 bp, or if they exhibited >10 sequencing errors. Next, we separated the reads of the template strand from the nontemplate strand by identifying the C-to-T conversion events at the repeats (identified as TGG *vs.* TTG if sense or antisense, respectively). Reads with no apparent conversions (43%) were excluded from further analysis since they could not be ascribed to any of the DNA strands. This subpopulation of molecules represents either paired DNA or a template strand paired with RNA (DNA:RNA hybrids), as would be expected in R-loop–forming regions. In addition, we excluded all molecules with unexplained deletions or with a CGG repeat number that deviated from 10 copies. Next, we clustered the reads within each group according to their bisulfite conversion patterns, which yielded two independent heatmaps (for the template strand and the nontemplate strand), representing the entire repertoire of molecules with DNA unpairing in that region (Figure 5A).

We thus identified distinct conversion designs that strongly support the systematic looping out of DNA strands at preferred locations, while avoiding data misinterpretation due to PCR duplicates. In the sense orientation, four distinct patterns could be identified, the majority (94%) of which covered the region between the TSS and the repeats (60 bp). More than half (66%) initiated at ~35 bp upstream to the TSS. Of these, the most common (55%) spread out into the 5′ portion of the repeats. Note that in no case was the full array of repeats (CGG) completely unpaired, thus supporting the preliminary data obtained by single-colony sequencing. In addition, a significant percentage (13%) of the DNA displacements extended further downstream and were systematically interrupted by double-strand segments.

In the antisense orientation, two distinct patterns could be identified, where the initiation and termination sites of extrusion were less flexible, and did not initiate upstream to the TSS. In the majority (60%), looping out was centered on the region that spans from the TSS to the first half of the repeats. In a considerable number of molecules, looping out was much more constrained and apparently interrupted, accumulating at the TSS and the repeats. In no case were the repeats (GCC) entirely unpaired, thus corresponding with what was observed for the nontemplate strand.

Altogether, the bisulfite strand-specific DNA labeling at/next to the repeats and upstream to the TSS strongly indicates that the DNA is commonly unpaired in this region. In searching for common/complementary designs between the strands, we

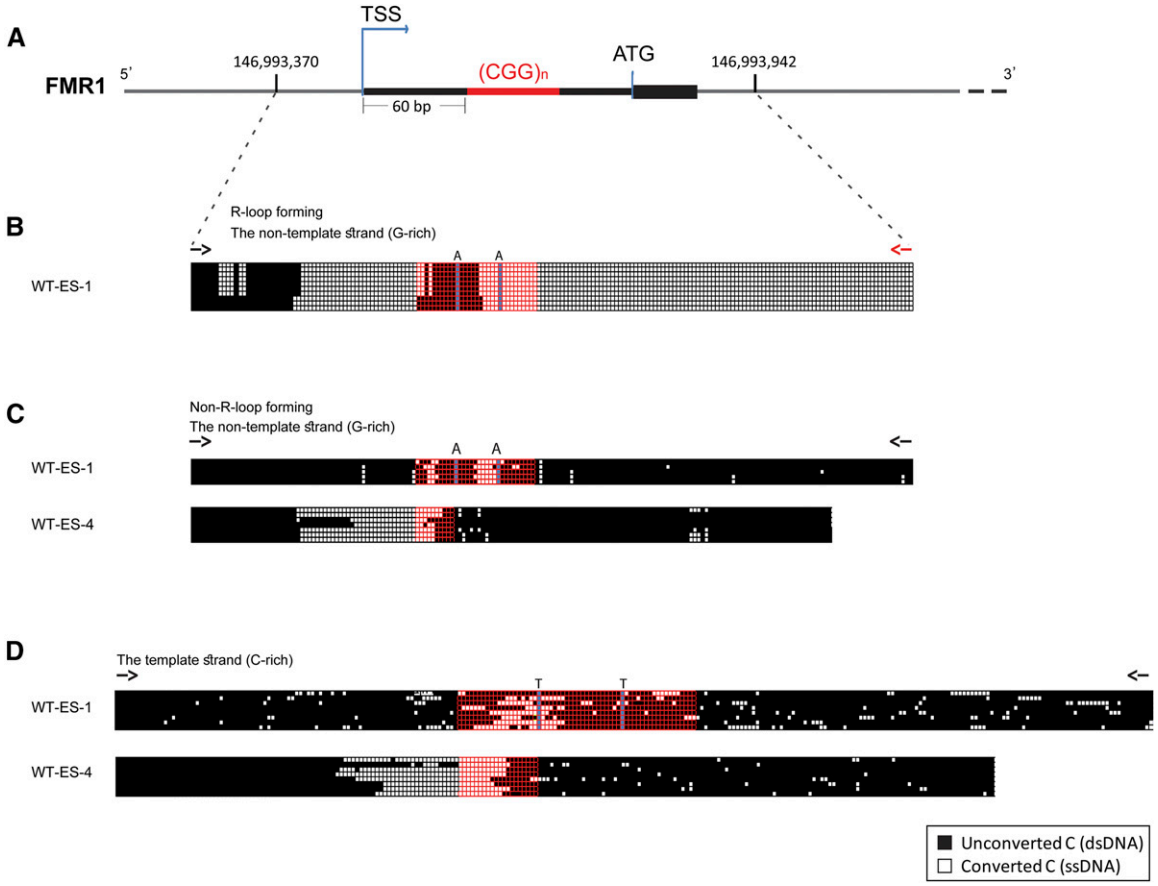


Figure 4 Bisulfite footprinting across the *FMR1* repeats in WT XY hESCs. (A) Schematic as in Figure 2B. (B) DNA bisulfite footprinting was carried out on the nontemplate strand (G-rich, sense) in WT XY hESC (WT-ES-1), in a region that spans the CGG repeats, using a pair of converted and unconverted primers. Black and red dashed arrows designate unconverted and converted primers, respectively. Each square represents a single cytosine site. Black squares represent bisulfite unconverted Cs (indicating double-strand DNA conformation) and white squares represent bisulfite converted Cs to Ts (indicating single-strand DNA conformation). Red outlined squares represent the repeat region. Blue squares represent AGG interruptions. WT-ES-1 carries 31 CGG repeats with two AGG interruptions. Note the double-strand DNA block at the 5′ half of the CGG repeats. (C) DNA bisulfite footprinting was carried out on the nontemplate strand (G-rich) in two WT XY hESC (WT-ES-1 and WT-ES-4, which carries 10 CGG repeats with no AGG interruptions) using two unconverted primers, enriching a population of molecules without R-loops. Each row represents a single DNA molecule and each square represents a single cytosine site. Black squares represent bisulfite unconverted C (indicating double-strand DNA conformation) and white squares represent bisulfite converted C to T (indicating single-strand DNA conformation). Note the DNA unwinding interruptions at and near the CGG repeats. (D) DNA bisulfite footprinting was carried out on the template strand (C-rich, antisense) in two different XY WT hESC lines, in a region that spans the CGG repeats, using a pair of unconverted primers. For simplicity, the template strand is presented from 3′ to 5′ (according to the nontemplate strand orientation). All symbols are as in Figure 2B. Note the single-strand DNA blocks (white) that are located within the double-strand DNA (black) across the repeats.

also point to uneven bisulfite strand-specific labeling, which is the hallmark of asymmetric DNA unpairing.

We posited that when hyper-methylation is induced and gene transcription is suppressed, DNA unpairing would be abolished, given the tight link between transcriptional activity and CGG instability in FXS cells (Wöhrle *et al.* 2001; Avitzour *et al.* 2014). To test this supposition, we looked for DNA unpairing events at the repeats in hESCs with a methylated allele. Since *FMR1* is prone to DNA methylation by X-inactivation in females, we carried out bisulfite footprinting under nondenatured conditions in an FXS XX hESC line that exhibits skewed X-inactivation with exclusive inactivation of the normal allele [see Figure S3 and Avitzour *et al.* (2014)]. Hence, skewed X inactivation specifically in this cell line provided an exceptional opportunity to investigate a wild-type yet methylated and transcriptionally inactive allele.

Based on the analysis of C-to-T conversions in non-CpG sites at the repeats and flanking regions, we parsed sequencing reads to template and nontemplate strands. Reads with no conversions (representing molecules that are entirely base-paired) were excluded from analysis. Next, we clustered the reads within each group and generated heatmaps according to bisulfite conversion patterns. The results showed a complete loss of ssDNA interruptions in both orientations (Figure 5B). These novel findings suggest a potential mechanistic link between local DNA unpairing and active transcription and/or open chromatin configuration in the *FMR1* gene.

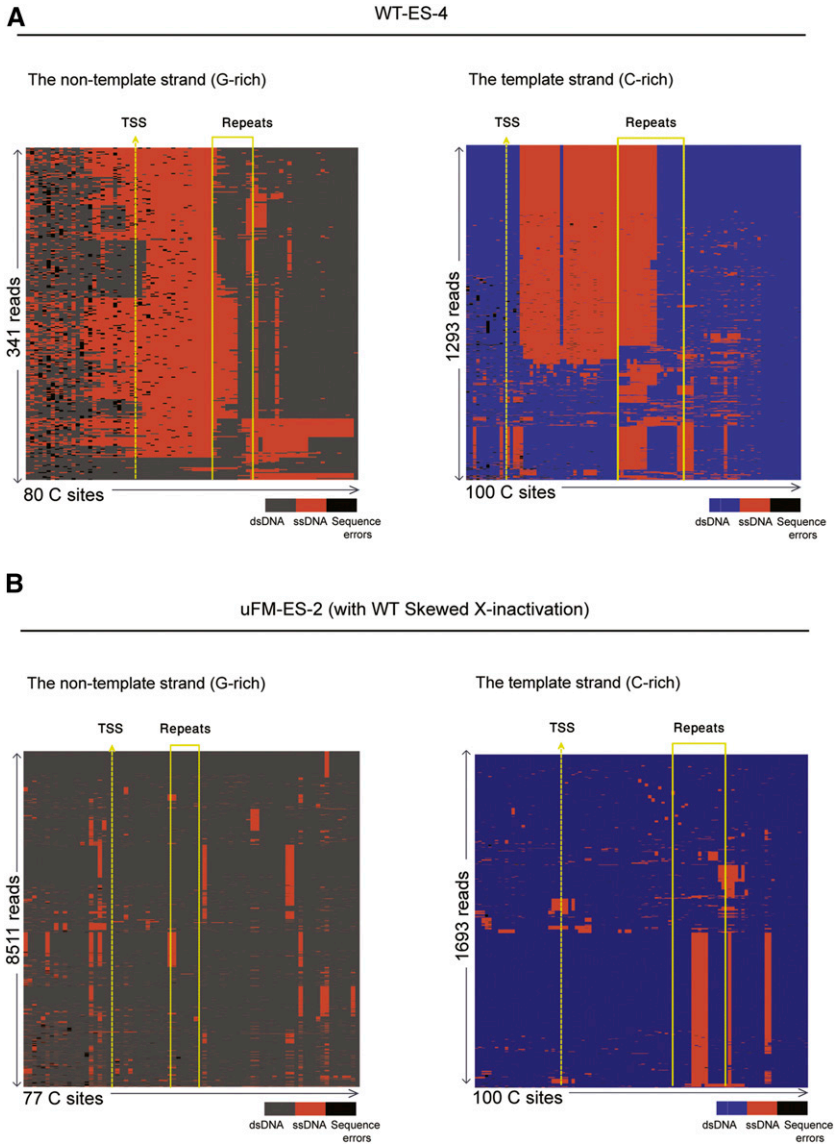


Figure 5 Bisulfite footprinting by deep-sequencing across the *FMR1* repeats in wild-type (WT) hESCs. (A) DNA bisulfite footprinting by deep-sequencing was carried out using unconverted primers in WT XY hESC (WT-ES-4). This was followed by a bioinformatic analysis which separated the reads into nontemplate (G-rich; left panel) and template (C-rich; right panel) strands, according to conversion patterns. Next, the reads were clustered into heatmaps. For simplicity, the template strand is presented in an opposite orientation (from 3′ to 5′ similar to the nontemplate strand orientation). The total read count appears on the *y*-axis. The length of the analyzed region is 250 bp with 80 C sites for the nontemplate and 100 C sites for the template strand. Dark gray and blue represent double-strand DNA (dsDNA) at the nontemplate and template strands, respectively, red represents single-strand DNA (ssDNA), and black represents sequencing errors. The TSS site and the repeats are designated with yellow lines. (B) DNA bisulfite footprinting by deep-sequencing was carried out using unconverted primers in a FXS XX hESC lines with an unmethylated full expansion (uFM) with skewed X-inactivation of the WT allele (uFM-ES-2), which allowed the selective amplification of a methylated WT allele. This was followed by a bioinformatic analysis, which separated the reads into nontemplate (G-rich; left panel) and template (C-rich; right panel) strands, according to conversion patterns.

DNA unpairing at/near the C9 repeats

Based on the data on the *FMR1* locus, we examined whether similar structures were regularly formed at the *C9orf72* locus. Although we failed to find bisulfite footprint G-rich R-loop–coupled repeats, we footprinted the C-rich DNA strand in two different wild-type hESC lines (WT-ES-1 and WT-ES-5) using a pair of unconverted primers. This provided evidence for similar interruptions by DNA unpairing at the CCCCGG repeats and downstream flanking sequence (Figure 6A). To better characterize this feature and determine how widespread it is, we repeated the analysis by bisulfite deep-sequencing under nondenatured conditions in four different hESC lines representing nonexpanded alleles comprised of two or five repeats (Figure 6B and Figure S4). Unexpectedly, no amplicons from the nontemplate strand were obtained. Nevertheless, using exactly the same approach as for *FMR1*, we identified two predominant patterns in the template strand reads: one that included ssDNA interruptions at the repeats and extends further downstream, terminating at an alternative TSS (TSS V2, a fragment of ∼46–83 bp, 23–48% molecules), and the other peaking around TSS V2 (∼29–72 bp, 52% molecules) (Figure 6B and Figure S4). Unconverted molecules were rarely observed (0.25–7%) and omitted from the analysis.

Finally, to explore whether restricted unpairing is a unique feature to the G-rich repeats in the *FMR1* and *C9of72* genes rather than a collective event of DNA melting by RNA Pol II, we carried out bisulfite colony footprinting at the TSS of two additional genes that are driven by CpG island promoters and identified as R-loop–forming regions but without repeats (*GAPDH* and *RPL13A*). Using the same approach as described above for the template (C-rich) and nontemplate (G-rich) DNA strands, we ruled out the possibility of recurrent DNA unpairing occurring at other unrelated loci in the genome (Figure S5).

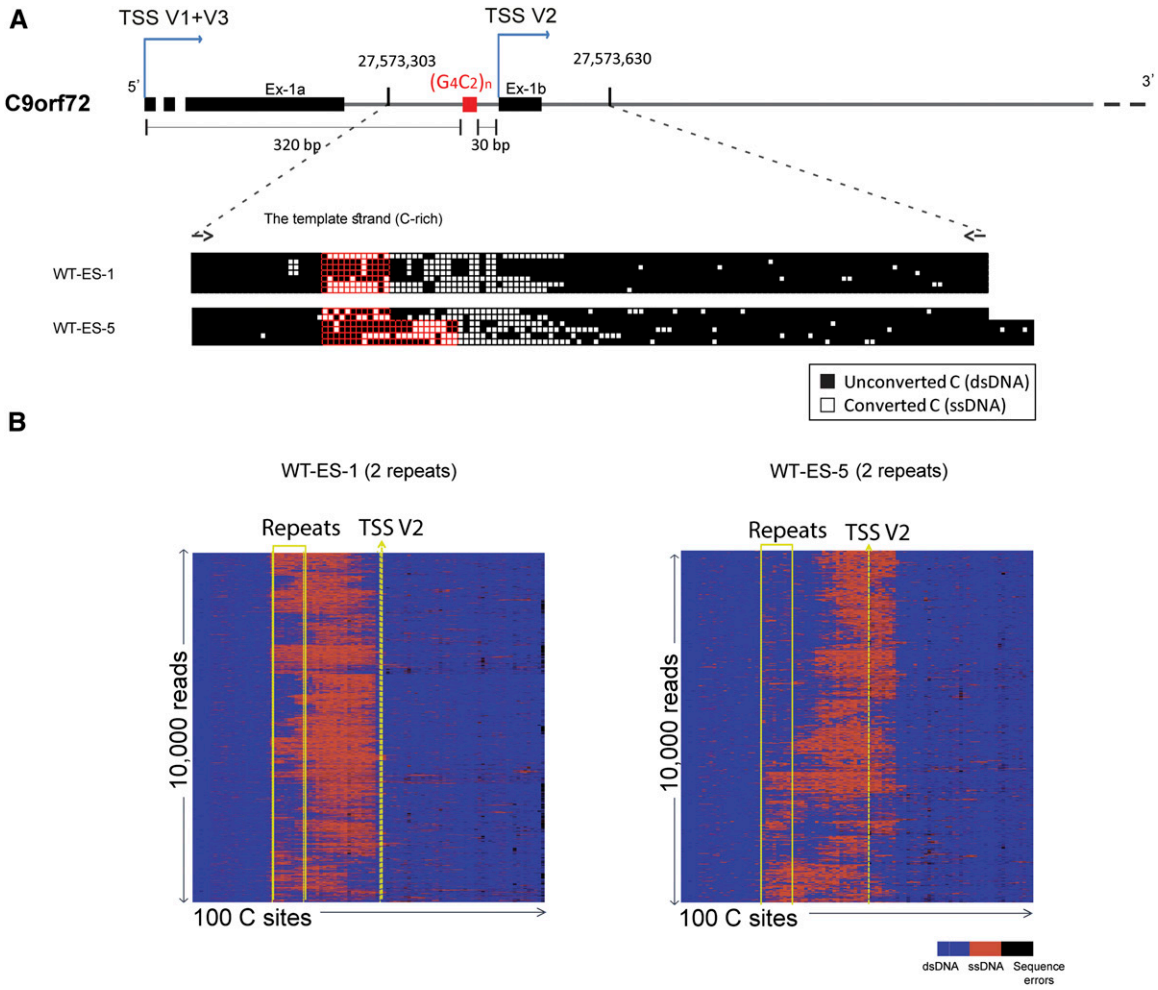


Figure 6 Bisulfite footprinting across the *C9orf72* repeats in wild-type (WT) hESCs alleles. (A) Schematic representing the 5′-UTR of *C9orf72* harboring the 5′-end of the transcript and two alternate TSS sites relative to the repeats (black). DNA colony bisulfite footprinting data obtained from the template strand (C-rich) using a pair of unconverted primers in two different XY WT hESC lines (WT-ES-1 and WT-ES-5) is presented. WT-ES-1 carries two GGGGCC repeats, whereas WT-ES-5 is a heterozygote for two and five GGGGCC repeats. For simplicity, the template strand is presented from 3′ to 5′ (according to the nontemplate strand orientation). Note the single-strand DNA (ssDNA) blocks (white) that are located within the double-strand DNA (dsDNA; black) across and near the repeats. Each row represents a single DNA molecule and each square represents a single cytosine site. Black squares represent bisulfite unconverted C (indicating dsDNA conformation) and white squares represent bisulfite converted C to T (indicating ssDNA conformation). (B) DNA bisulfite footprinting by deep-sequencing was carried out using unconverted primers in two different WT hESC lines (WT-ES-1 and WT-ES-5). This was followed by a bioinformatic analysis, which revealed that all reads were derived exclusively from the template strand (C-rich), according to conversion patterns. Next, 10,000 reads were randomly selected for clustering into heatmaps. The length of the analyzed region was 212 bp with 100 C sites for the template strand.

Altogether, we provide firm evidence for the comprehensive looping-out of DNA at the *FMR1* (G- and C-rich templates) and *C9orf72* (C-rich template) loci, thus supporting the supposition that local DNA unpairing is a typical feature of G-rich repetitive regions in actively transcribed genes that are prone to instability.

Discussion

Evidence emerging from diverse systems, including human cells, suggests that R-loop coupled transcription can lead to repeat instability by promoting ssDNA displacements, a potent source of non-B DNA secondary structures (Panigrahi *et al.* 2005; Grabczyk *et al.* 2007; Lin *et al.* 2010; Salinas-Rios *et al.* 2011; Reddy *et al.* 2014; Slean *et al.* 2016). In the current study we characterized the formation of R-loops and finely mapped ssDNA displacements near and across the G-rich repeats at the *FMR1* and *C9orf72* genes. This should lead to a better understanding of the ways in which these regions are predisposed to repeat instability. The argument that R-loops mediate repeat instability through the induction of ssDNA displacements in these regions is particularly appealing since in both genes, the repeats are transcribed and positioned next to CpG island promoters, providing preferred sites for R-loop accumulation (Ginno *et al.* 2012).

In this study, we used hESCs with wild-type and expanded alleles (Eiges *et al.* 2007; Avitzour *et al.* 2014; Cohen-Hadad *et al.* 2016) in the *FMR1* or *C9orf72* genes, to characterize the

ssDNA displacements that are frequently formed in those gene regions. After confirming R-loop enrichments in both loci by DRIP analysis (Figure 1) (Colak *et al.* 2014; Groh *et al.* 2014; Loomis *et al.* 2014; Kumari and Usdin 2016; Esanov *et al.* 2017), we finely mapped their boundaries at near-nucleotide resolution. Using bisulfite footprinting, we identified the initiation (5′-end) and termination (3′-end) sites of the three-way junctions in *FMR1* and *C9orf72* (Figure 2 and Figure 3). We showed that DNA displacements on the G-rich strand occur at preferable sites and are found in wild-type and expanded alleles. This, together with the DRIP data and the composition of the DNA sequence immediately downstream to the TSS (Chen *et al.* 2017), strongly support the notion that these three-way junctions represent the ends of the R-loop. This leads to the inference that the length of the R-loop in wild-type alleles is at least ~550 bp in *FMR1* and ~800 bp in *C9orf72*. This is rather unusual for promoter-associated R-loops since the majority are considerably shorter and terminate at the first exon-intron junction (Dumelie and Jaffrey 2017). Furthermore, our results are the first to precisely map the 3′-end of the R-loop in *FMR1* and to identify the initiation and termination sites of the R-loops in *C9orf72*. In addition, they are in complete agreement with an earlier study that identified the 5′-end of the R-loop in *FMR1* (Loomis *et al.* 2014).

Interestingly, when we bisulfite-footprinted a region that initiates from the TSS and extends past the CGGs in *FMR1*, we noticed that the ssDNA displacements, when coupled with R-loops, were consistently interrupted by short invariable blocks of dsDNA segments at the 5′ half of the repeats (Figure 4B). These fragments were only detected in FXS unaffected cells (<200 CGGs) due to technical limitations in amplifying the repetitive region. Nevertheless, this result extends findings reported by Loomis and colleagues, who documented the existence of similar interruptions in fibroblast cell cultures with wild-type or premutation (55 > CGGs > 200) alleles (Loomis *et al.* 2014). Although the existence of multiple R-loops cannot be completely ruled out, it is less likely to be the case. This is because cell-free studies, corroborated by chromatin immunoprecipitation–sequencing data, provide convincing evidence for the formation of complex secondary structures by long tracts of G-rich microsatellite repeats (Fry and Loeb 1994; Moore *et al.* 1999; Fukuda *et al.* 2005; Chambers *et al.* 2015; Hänsel-Hertsch *et al.* 2016).

By applying bisulfite footprinting using unconverted primers, we identified recurrent events of restricted unwinding in the DNA across and near the repeats, as clearly illustrated by the heatmaps (Figure 5, Figure 6, and Figure S4). This resulted in distinct patterns ascribed to the G-rich nontemplate (*FMR1*) and C-rich template strand (*FMR1* and *C9orf72*), the majority of which cover the region that spans between a nearby TSS and the repeats. We speculate that the propensity of these regions to recurrently unwind makes them favorable sites, or hotspots, for DNA unwinding due to the G-richness and the repetitive nature of their sequence.

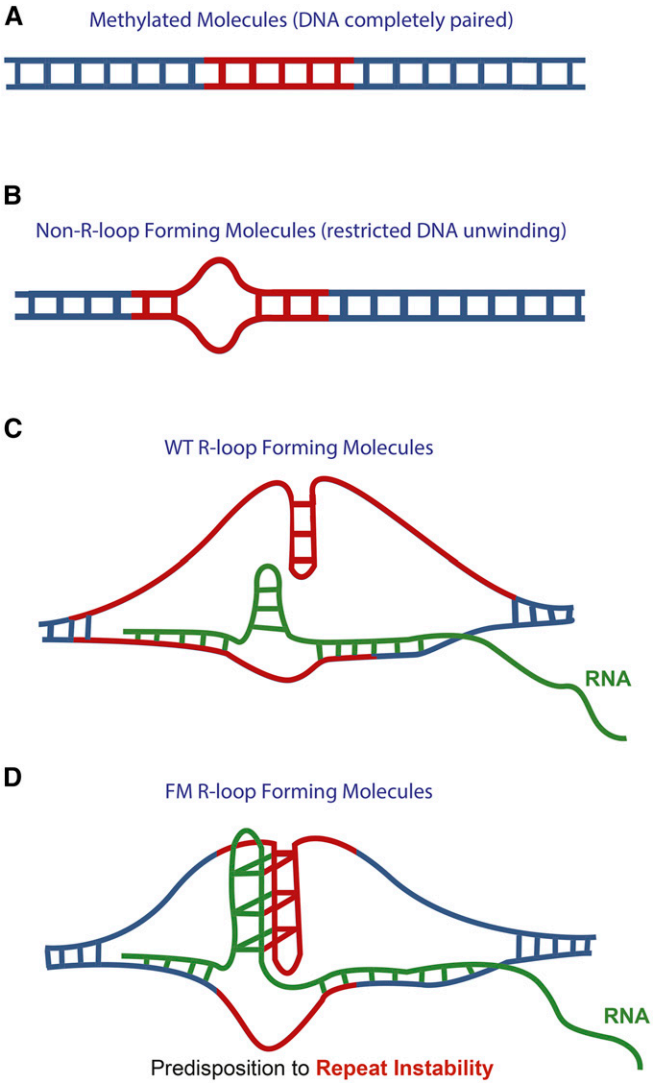


Figure 7 Proposed model for the formation of noncanonical structures by the G/C-rich repeats at the *FMR1* and *C9orf72* loci. Four potential configurations can be formed at the repeats in *FMR1* and *C9orf72* loci: (A) when the gene is transcriptionally inactive and completely DNA:DNA paired; (B) when the chromatin is in a transcriptionally competent state, there are recurrent events of restricted DNA unwinding at/near the repeats; (C) in normal range alleles, when R-loops are formed, DNA unwinding becomes much more extensive and results in the formation of noncanonical (non-B form) double-strand structure(s) by the nontemplate DNA strand (illustrated as red hairpin); and (D) long CGG/GGGGCC in the pathologic range, when coupled with R-loops, generate more difficult to process structures than their shorter counterparts. The newly synthesized lengthy RNA can assemble into similar secondary structures as the nontemplate DNA (illustrated as green hairpin) and jointly form a complex RNA:DNA hybrid (Zheng *et al.* 2013; Zhang *et al.* 2014). Such G-rich hybrid structures are expected to be particularly stable and difficult for the cell to resolve, thus providing a potent source of repeat instability. Blue, red, and green lines designate the DNA (paired and unpaired), the CGG/GGGGCC repeats on the nontemplate/template strand, and the newly synthesized RNA molecules, respectively. FM, full mutation; WT, wild type.

Strikingly, when the *FMR1* gene was hyper-methylated and transcriptionally inactive (in the FXS XX hESC line that exhibits skewed X-inactivation), local unpairing was

abolished (Figure 5B). Since DNA unwinding occurs near the TSS of *FMR1*, this encouraged us to posit that this reflects DNA melting by RNA Pol II pausing. In fact, this is consistent with the transcription-induced repeat instability model, which suggests that when RNA Pol II encounters unusual structures generated by the repeats, it pauses. This activates the transcription coupled repair machinery (Lin and Wilson 2007; Lin *et al.* 2009; Sollier *et al.* 2014), ultimately leading to the gain/loss of repeats. On the other hand, asymmetric bisulfite labeling of the complementary strands, as clearly illustrated by the heatmaps, argues against this possibility. The alternative mechanisms that promote repeat instability and that are involved in the processing of unconventional DNA structures include incorrect recruitment of mismatch repair proteins and activation of the base excision repair pathway (Du *et al.* 2012; Gannon *et al.* 2012; Halabi *et al.* 2012; Lokanga *et al.* 2014, 2015; Zhao *et al.* 2015). Other potential mechanisms for repeat instability which rely on DNA replication include DNA polymerase stalling and fork collapse, which result in error-prone repair, and DNA polymerase strand-slippage during DNA synthesis (Usdin and Woodford 1995; Voineagu *et al.* 2009; Gerhardt *et al.* 2014, 2016; Kononenko *et al.* 2018). Regardless of whether the mechanism is repair or replication, our results are consistent with earlier studies on the induction of ssDNA displacements by long microsatellite repeat tracts (Pearson and Sinden 1996; Pearson *et al.* 1998b, 2002; Tam *et al.* 2003; Panigrahi *et al.* 2005, 2010; Axford *et al.* 2013; Slean *et al.* 2013, 2016; Kononenko *et al.* 2018).

To summarize, we provide evidence that the G-rich repeats in *FMR1* and *C9orf72* tend to be unpaired when the chromatin is in a transcriptionally competent state. In this condition, when R-loops form, DNA unpairing becomes more extensive, providing an opportunity for the nontemplate DNA strand to adopt a noncanonical (non-B form) double-strand structure(s). It is anticipated that long CGG/GGGGCC tracts should generate more difficult to process structures than their shorter counterparts (see Figure 7 for a proposed model). Furthermore, it is in theory possible that newly synthesized lengthy RNA will assemble into similar secondary structures and interact with the nontemplate DNA to jointly form a RNA:DNA hybrid (Zheng *et al.* 2013; Zhang *et al.* 2014). Such G-rich hybrid structures are expected to be particularly stable and difficult for the cell to resolve. Once identified, these mutant non-B structures are likely to be associated with the induction of dsDNA breaks or replication irregularities to functionally associate them with repeat instability in the *FMR1/C9orf72* loci.

Acknowledgments

We thank the families who donated the FXS and C9/ALS embryos for hESC line derivation. We would also like to thank Amir Eden for fruitful discussions and David Zeevi for critically reading the manuscript, Motti Peretz for assistance with the graphic design, Anna Typsin for the illustrations, and Clinton E. Leysath for the S9.6 monoclonal antibody. We thank the Genomic Applications Laboratory, The Core Research Facility, Faculty of Medicine – Ein Kerem, The Hebrew University of Jerusalem, Israel for the next-generation sequencing. This research was supported by the Israel Science Foundation (grant 1480/15 to R.E.) and the Legacy Heritage Biomedical Program of the Israel Science Foundation (grant 1260/16 to R.E.). We declare no conflicts of interest.

Literature Cited

Aguilera, A., and T. Garcia-Muse, 2012 R loops: from transcription byproducts to threats to genome stability. Mol. Cell 46: 115–124. https://doi.org/10.1016/j.molcel.2012.04.009

Avitzour, M., H. Mor-Shaked, S. Yanovsky-Dagan, S. Aharoni, G. Altarescu *et al.*, 2014 FMR1 epigenetic silencing commonly occurs in undifferentiated fragile X-affected embryonic stem cells. Stem Cell Reports 3: 699–706. https://doi.org/10.1016/j.stemcr.2014.09.001

Axford, M. M., Y. H. Wang, M. Nakamori, M. Zannis-Hadjopoulos, C. A. Thornton *et al.*, 2013 Detection of slipped-DNAs at the trinucleotide repeats of the myotonic dystrophy type I disease locus in patient tissues. PLoS Genet. 9: e1003866. https://doi.org/10.1371/journal.pgen.1003866

Boque-Sastre, R., M. Soler, and S. Guil, 2017 Detection and characterization of R loop structures. Methods Mol. Biol. 1543: 231–242. https://doi.org/10.1007/978-1-4939-6716-2_13

Chambers, V. S., G. Marsico, J. M. Boutell, M. Di Antonio, G. P. Smith *et al.*, 2015 High-throughput sequencing of DNA G-quadruplex structures in the human genome. Nat. Biotechnol. 33: 877–881. https://doi.org/10.1038/nbt.3295

Chen, L., J. Y. Chen, X. Zhang, Y. Gu, R. Xiao *et al.*, 2017 R-ChIP using inactive RNase H reveals dynamic coupling of R-loops with transcriptional pausing at gene promoters. Mol. Cell 68: 745–757.e5. https://doi.org/10.1016/j.molcel.2017.10.008

Cleary, J. D., K. Nichol, Y. H. Wang, and C. E. Pearson, 2002 Evidence of cis-acting factors in replication-mediated trinucleotide repeat instability in primate cells. Nat. Genet. 31: 37–46. https://doi.org/10.1038/ng870

Cohen-Hadad, Y., G. Altarescu, T. Eldar-Geva, E. Levi-Lahad, M. Zhang *et al.*, 2016 Marked differences in C9orf72 methylation status and isoform expression between C9/ALS human embryonic and induced pluripotent stem cells. Stem Cell Reports 7: 927–940. https://doi.org/10.1016/j.stemcr.2016.09.011

Colak, D., N. Zaninovic, M. S. Cohen, Z. Rosenwaks, W. Y. Yang *et al.*, 2014 Promoter-bound trinucleotide repeat mRNA drives epigenetic silencing in fragile X syndrome. Science 343: 1002–1005. https://doi.org/10.1126/science.1245831

DeJesus-Hernandez, M., I. R. Mackenzie, B. F. Boeve, A. L. Boxer, M. Baker *et al.*, 2011 Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron 72: 245–256. https://doi.org/10.1016/j.neuron.2011.09.011

Dols-Icardo, O., A. Garcia-Redondo, R. Rojas-Garcia, R. Sanchez-Valle, A. Noguera *et al.*, 2014 Characterization of the repeat expansion size in C9orf72 in amyotrophic lateral sclerosis and frontotemporal dementia. Hum. Mol. Genet. 23: 749–754. https://doi.org/10.1093/hmg/ddt460

Du, J., E. Campau, E. Soragni, S. Ku, J. W. Puckett *et al.*, 2012 Role of mismatch repair enzymes in GAA·TTC triplet-repeat expansion in Friedreich ataxia induced pluripotent stem cells. J. Biol. Chem. 287: 29861–29872. https://doi.org/10.1074/jbc.M112.391961

Dumelie, J. G., and S. R. Jaffrey, 2017 Defining the location of promoter-associated R-loops at near-nucleotide resolution using

bisDRIP-seq. eLife 6: e28306. https://doi.org/10.7554/eLife.28306

Eiges, R., A. Urbach, M. Malcov, T. Frumkin, T. Schwartz *et al.*, 2007 Developmental study of fragile X syndrome using human embryonic stem cells derived from preimplantation genetically diagnosed embryos. Cell Stem Cell 1: 568–577. https://doi.org/10.1016/j.stem.2007.09.001

Esanov, R., G. T. Cabrera, N. S. Andrade, T. F. Gendron, R. H. Brown *et al.*, 2017 A C9ORF72 BAC mouse model recapitulates key epigenetic perturbations of ALS/FTD. Mol. Neurodegener. 12: 46. https://doi.org/10.1186/s13024-017-0185-9

Fry, M., and L. A. Loeb, 1994 The fragile X syndrome d(CGG)n nucleotide repeats form a stable tetrahelical structure. Proc. Natl. Acad. Sci. USA 91: 4950–4954. https://doi.org/10.1073/pnas.91.11.4950

Fukuda, H., M. Katahira, E. Tanaka, Y. Enokizono, N. Tsuchiya *et al.*, 2005 Unfolding of higher DNA structures formed by the d(CGG) triplet repeat by UP1 protein. Genes Cells 10: 953–962. https://doi.org/10.1111/j.1365-2443.2005.00896.x

Gannon, A. M., A. Frizzell, E. Healy, and R. S. Lahue, 2012 MutSβ and histone deacetylase complexes promote expansions of trinucleotide repeats in human cells. Nucleic Acids Res. 40: 10324–10333. https://doi.org/10.1093/nar/gks810

Gatchel, J. R., and H. Y. Zoghbi, 2005 Diseases of unstable repeat expansion: mechanisms and common principles. Nat. Rev. Genet. 6: 743–755. https://doi.org/10.1038/nrg1691

Gerhardt, J., N. Zaninovic, Q. Zhan, A. Madireddy, S. L. Nolin *et al.*, 2014 Cis-acting DNA sequence at a replication origin promotes repeat expansion to fragile X full mutation. J. Cell Biol. 206: 599–607. https://doi.org/10.1083/jcb.201404157

Gerhardt, J., A. D. Bhalla, J. S. Butler, J. W. Puckett, P. B. Dervan *et al.*, 2016 Stalled DNA replication forks at the endogenous GAA repeats drive repeat expansion in Friedreich's ataxia cells. Cell Rep. 16: 1218–1227. https://doi.org/10.1016/j.celrep.2016.06.075

Ginno, P. A., P. L. Lott, H. C. Christensen, I. Korf, and F. Chedin, 2012 R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol. Cell 45: 814–825. https://doi.org/10.1016/j.molcel.2012.01.017

Grabczyk, E., M. Mancuso, and M. C. Sammarco, 2007 A persistent RNA.DNA hybrid formed by transcription of the Friedreich ataxia triplet repeat in live bacteria, and by T7 RNAP in vitro. Nucleic Acids Res. 35: 5351–5359. https://doi.org/10.1093/nar/gkm589

Gray, L. T., A. C. Vallur, J. Eddy, and N. Maizels, 2014 G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. Nat. Chem. Biol. 10: 313–318. https://doi.org/10.1038/nchembio.1475

Groh, M., M. M. Lufino, R. Wade-Martins, and N. Gromak, 2014 R-loops associated with triplet repeat expansions promote gene silencing in Friedreich ataxia and fragile X syndrome. PLoS Genet. 10: e1004318. https://doi.org/10.1371/journal.pgen.1004318

Halabi, A., S. Ditch, J. Wang, and E. Grabczyk, 2012 DNA mismatch repair complex MutSβ promotes GAA·TTC repeat expansion in human cells. J. Biol. Chem. 287: 29958–29967. https://doi.org/10.1074/jbc.M112.356758

Hänsel-Hertsch, R., D. Beraldi, S. V. Lensing, G. Marsico, K. Zyner *et al.*, 2016 G-quadruplex structures mark human regulatory chromatin. Nat. Genet. 48: 1267–1272. https://doi.org/10.1038/ng.3662

Jenjaroenpun, P., T. Wongsurawat, S. P. Yenamandra, and V. A. Kuznetsov, 2015 QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. Nucleic Acids Res. 43: W527–W534. https://doi.org/10.1093/nar/gkv344

Jenjaroenpun, P., T. Wongsurawat, S. Sutheeworapong, and V. A. Kuznetsov, 2017 R-loopDB: a database for R-loop forming sequences (RLFS) and R-loops. Nucleic Acids Res. 45: D119–D127. https://doi.org/10.1093/nar/gkw1054

Jonkers, I., and J. T. Lis, 2015 Getting up to speed with transcription elongation by RNA polymerase II. Nat. Rev. Mol. Cell Biol. 16: 167–177. https://doi.org/10.1038/nrm3953

Kononenko, A. V., T. Ebersole, K. M. Vasquez, and S. M. Mirkin, 2018 Mechanisms of genetic instability caused by (CGG)n repeats in an experimental mammalian system. Nat. Struct. Mol. Biol. 25: 669–676. https://doi.org/10.1038/s41594-018-0094-9

Kumari, D., and K. Usdin, 2016 Sustained expression of FMR1 mRNA from reactivated fragile X syndrome alleles after treatment with small molecules that prevent trimethylation of H3K27. Hum. Mol. Genet. 25: 3689–3698. https://doi.org/10.1093/hmg/ddw215

Lin, Y., and J. H. Wilson, 2007 Transcription-induced CAG repeat contraction in human cells is mediated in part by transcription-coupled nucleotide excision repair. Mol. Cell. Biol. 27: 6209–6217. https://doi.org/10.1128/MCB.00739-07

Lin, Y., L. Hubert, and J. H. Wilson, 2009 Transcription destabilizes triplet repeats. Mol. Carcinog. 48: 350–361. https://doi.org/10.1002/mc.20488

Lin, Y., S. Y. Dent, J. H. Wilson, R. D. Wells, and M. Napierala, 2010 R loops stimulate genetic instability of CTG.CAG repeats. Proc. Natl. Acad. Sci. USA 107: 692–697. https://doi.org/10.1073/pnas.0909740107

Lokanga, R. A., X. N. Zhao, and K. Usdin, 2014 The mismatch repair protein MSH2 is rate limiting for repeat expansion in a fragile X premutation mouse model. Hum. Mutat. 35: 129–136. https://doi.org/10.1002/humu.22464

Lokanga, R. A., A. G. Senejani, J. B. Sweasy, and K. Usdin, 2015 Heterozygosity for a hypomorphic Polβ mutation reduces the expansion frequency in a mouse model of the Fragile X-related disorders. PLoS Genet. 11: e1005181. https://doi.org/10.1371/journal.pgen.1005181

Loomis, E. W., J. S. Eid, P. Peluso, J. Yin, L. Hickey *et al.*, 2013 Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res. 23: 121–128. https://doi.org/10.1101/gr.141705.112

Loomis, E. W., L. A. Sanz, F. Chedin, and P. J. Hagerman, 2014 Transcription-associated R-loop formation across the human FMR1 CGG-repeat region. PLoS Genet. 10: e1004294. https://doi.org/10.1371/journal.pgen.1004294

Mirkin, S. M., 2007 Expandable DNA repeats and human disease. Nature 447: 932–940. https://doi.org/10.1038/nature05977

Moore, H., P. W. Greenwell, C. P. Liu, N. Arnheim, and T. D. Petes, 1999 Triplet repeats form secondary structures that escape DNA repair in yeast. Proc. Natl. Acad. Sci. USA 96: 1504–1509. https://doi.org/10.1073/pnas.96.4.1504

Nichol Edamura, K., M. R. Leonard, and C. E. Pearson, 2005 Role of replication and CpG methylation in fragile X syndrome CGG deletions in primate cells. Am. J. Hum. Genet. 76: 302–311. https://doi.org/10.1086/427928

Oberle, I., F. Rousseau, D. Heitz, C. Kretz, D. Devys *et al.*, 1991 Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. Science 252: 1097–1102. https://doi.org/10.1126/science.252.5009.1097

Panigrahi, G. B., R. Lau, S. E. Montgomery, M. R. Leonard, and C. E. Pearson, 2005 Slipped (CTG)*(CAG) repeats can be correctly repaired, escape repair or undergo error-prone repair. Nat. Struct. Mol. Biol. 12: 654–662. https://doi.org/10.1038/nsmb959

Panigrahi, G. B., M. M. Slean, J. P. Simard, O. Gileadi, and C. E. Pearson, 2010 Isolated short CTG/CAG DNA slip-outs are repaired efficiently by hMutSbeta, but clustered slip-outs are poorly repaired. Proc. Natl. Acad. Sci. USA 107: 12593–12598. https://doi.org/10.1073/pnas.0909087107

Panigrahi, G. B., M. M. Slean, J. P. Simard, and C. E. Pearson, 2012 Human mismatch repair protein hMutLα is required to

repair short slipped-DNAs of trinucleotide repeats. J. Biol. Chem. 287: 41844–41850. https://doi.org/10.1074/jbc.M112.420398

Pearson, C. E., and R. R. Sinden, 1996 Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. Biochemistry 35: 5041–5053. https://doi.org/10.1021/bi9601013

Pearson, C. E., E. E. Eichler, D. Lorenzetti, S. F. Kramer, H. Y. Zoghbi *et al.*, 1998a Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. Biochemistry 37: 2701–2708. https://doi.org/10.1021/bi972546c

Pearson, C. E., Y. H. Wang, J. D. Griffith, and R. R. Sinden, 1998b Structural analysis of slipped-strand DNA (S-DNA) formed in (CTG)n. (CAG)n repeats from the myotonic dystrophy locus. Nucleic Acids Res. 26: 816–823. https://doi.org/10.1093/nar/26.3.816

Pearson, C. E., M. Tam, Y. H. Wang, S. E. Montgomery, A. C. Dar *et al.*, 2002 Slipped-strand DNAs formed by long (CAG)*(CTG) repeats: slipped-out repeats and slip-out junctions. Nucleic Acids Res. 30: 4534–4547. https://doi.org/10.1093/nar/gkf572

Pearson, C. E., K. Nichol Edamura, and J. D. Cleary, 2005 Repeat instability: mechanisms of dynamic mutations. Nat. Rev. Genet. 6: 729–742. https://doi.org/10.1038/nrg1689

Reddy, K., M. H. Schmidt, J. M. Geist, N. P. Thakkar, G. B. Panigrahi *et al.*, 2014 Processing of double-R-loops in (CAG).(CTG) and C9orf72 (GGGGCC).(GGCCCC) repeats causes instability. Nucleic Acids Res. 42: 10473–10487. https://doi.org/10.1093/nar/gku658

Salinas-Rios, V., B. P. Belotserkovskii, and P. C. Hanawalt, 2011 DNA slip-outs cause RNA polymerase II arrest in vitro: potential implications for genetic instability. Nucleic Acids Res. 39: 7444–7454. https://doi.org/10.1093/nar/gkr429

Samadashwily, G. M., G. Raca, and S. M. Mirkin, 1997 Trinucleotide repeats affect DNA replication in vivo. Nat. Genet. 17: 298–304. https://doi.org/10.1038/ng1197-298

Slean, M. M., K. Reddy, B. Wu, K. Nichol Edamura, M. Kekis *et al.*, 2013 Interconverting conformations of slipped-DNA junctions formed by trinucleotide repeats affect repair outcome. Biochemistry 52: 773–785. https://doi.org/10.1021/bi301369b

Slean, M. M., G. B. Panigrahi, A. L. Castel, A. B. Pearson, A. E. Tomkinson *et al.*, 2016 Absence of MutSβ leads to the formation of slipped-DNA for CTG/CAG contractions at primate replication forks. DNA Repair (Amst.) 42: 107–118. https://doi.org/10.1016/j.dnarep.2016.04.002

Sollier, J., C. T. Stork, M. L. García-Rubio, R. D. Paulsen, A. Aguilera *et al.*, 2014 Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. Mol. Cell 56: 777–785. https://doi.org/10.1016/j.molcel.2014.10.020

Su, X. A., and C. H. Freudenreich, 2017 Cytosine deamination and base excision repair cause R-loop-induced CAG repeat fragility and instability in. Proc. Natl. Acad. Sci. USA 114: E8392–E8401. https://doi.org/10.1073/pnas.1711283114

Tam, M., S. Erin Montgomery, M. Kekis, B. D. Stollar, G. B. Price *et al.*, 2003 Slipped (CTG).(CAG) repeats of the myotonic dystrophy locus: surface probing with anti-DNA antibodies. J. Mol. Biol. 332: 585–600. https://doi.org/10.1016/S0022-2836(03)00880-5

Tassone, F., A. Beilina, C. Carosi, S. Albertosi, C. Bagni *et al.*, 2007 Elevated FMR1 mRNA in premutation carriers is due to increased transcription. RNA 13: 555–562. https://doi.org/10.1261/rna.280807

Usdin, K., and K. J. Woodford, 1995 CGG repeats associated with DNA instability and chromosome fragility form structures that block DNA synthesis in vitro. Nucleic Acids Res. 23: 4202–4209. https://doi.org/10.1093/nar/23.20.4202

Verkerk, A. J., M. Pieretti, J. S. Sutcliffe, Y. H. Fu, D. P. Kuhl *et al.*, 1991 Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65: 905–914. https://doi.org/10.1016/0092-8674(91)90397-H

Voineagu, I., C. F. Surka, A. A. Shishkin, M. M. Krasilnikova, and S. M. Mirkin, 2009 Replisome stalling and stabilization at CGG repeats, which are responsible for chromosomal fragility. Nat. Struct. Mol. Biol. 16: 226–228. https://doi.org/10.1038/nsmb.1527

Wöhrle, D., U. Salat, H. Hameister, W. Vogel, and P. Steinbach, 2001 Demethylation, reactivation, and destabilization of human fragile X full-mutation alleles in mouse embryocarcinoma cells. Am. J. Hum. Genet. 69: 504–515. https://doi.org/10.1086/322739

Yu, K., F. Chedin, C. L. Hsieh, T. E. Wilson, and M. R. Lieber, 2003 R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. Nat. Immunol. 4: 442–451. https://doi.org/10.1038/ni919

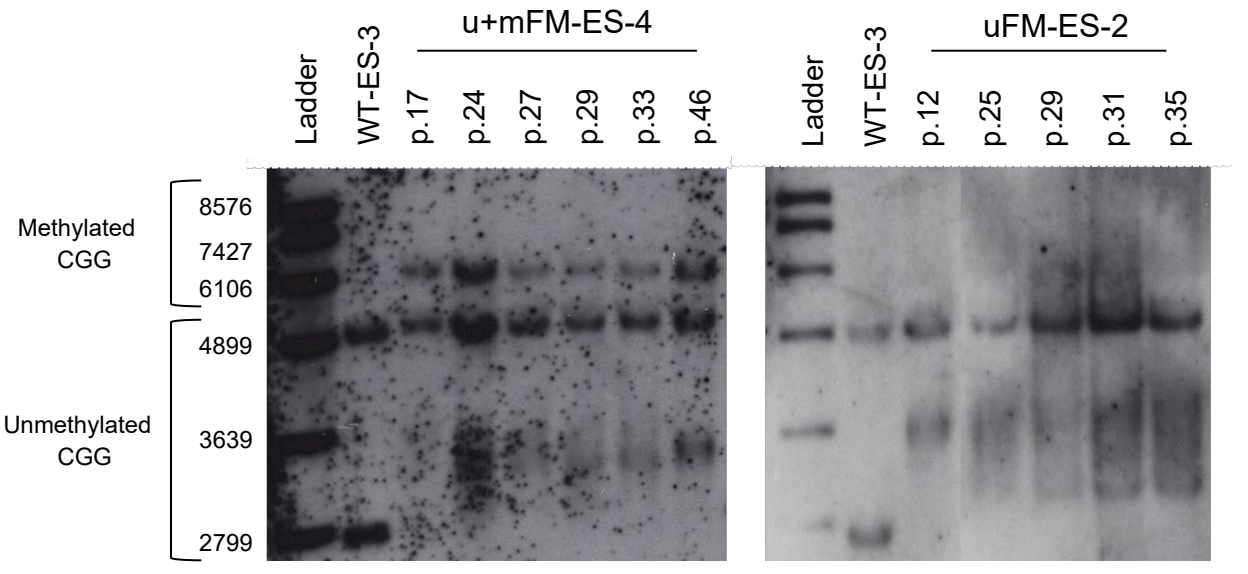
Zhang, J. Y., K. W. Zheng, S. Xiao, Y. H. Hao, and Z. Tan, 2014 Mechanism and manipulation of DNA:RNA hybrid G-quadruplex formation in transcription of G-rich DNA. J. Am. Chem. Soc. 136: 1381–1390. https://doi.org/10.1021/ja4085572

Zhao, X. N., D. Kumari, S. Gupta, D. Wu, M. Evanitsky *et al.*, 2015 Mutsβ generates both expansions and contractions in a mouse model of the Fragile X-associated disorders. Hum. Mol. Genet. 24: 7087–7096. https://doi.org/10.1093/hmg/ddv408

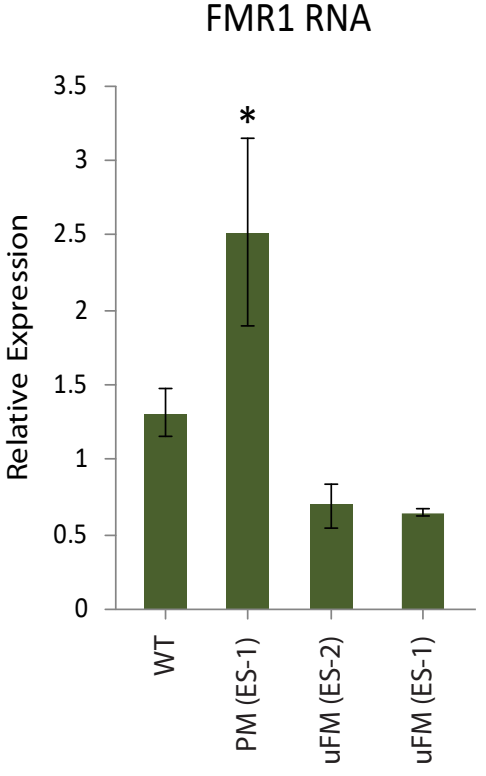
Zheng, K. W., S. Xiao, J. Q. Liu, J. Y. Zhang, Y. H. Hao *et al.*, 2013 Co-transcriptional formation of DNA:RNA hybrid G-quadruplex and potential function as constitutional cis element for transcription control. Nucleic Acids Res. 41: 5533–5541. https://doi.org/10.1093/nar/gkt264

Communicating editor: S. Sharan

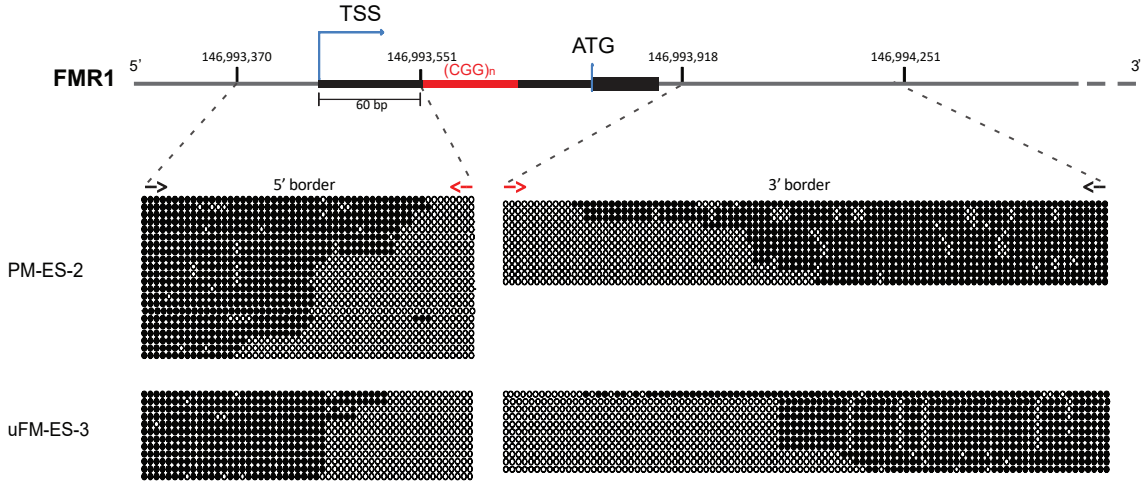
A


u+mFM-ES-4
uFM-ES-2
Ladder
WT-ES-3
p.17
p.24
p.27
p.29
p.33
p.46
Ladder
WT-ES-3
p.12
p.25
p.29
p.31
p.35
Methylated CGG
Unmethylated CGG
8576
7427
6106
4899
3639
2799

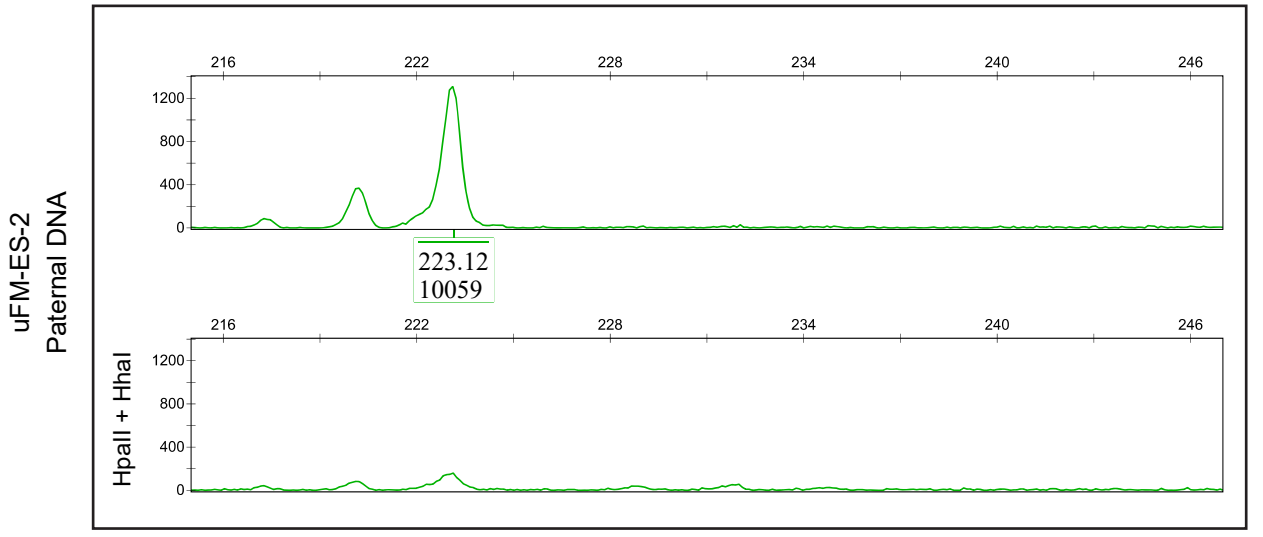
B

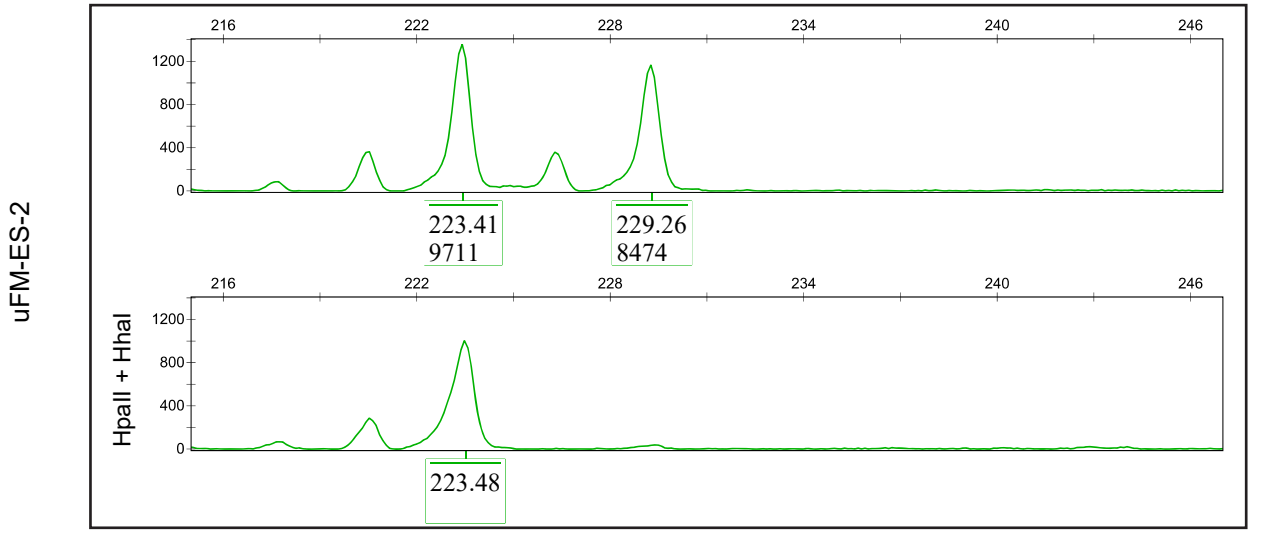

FMR1 RNA
Relative Expression
3.5
3
2.5
2
1.5
1
0.5
0
*
WT
PM (ES-1)
uFM (ES-2)
uFM (ES-1)

Sup 2

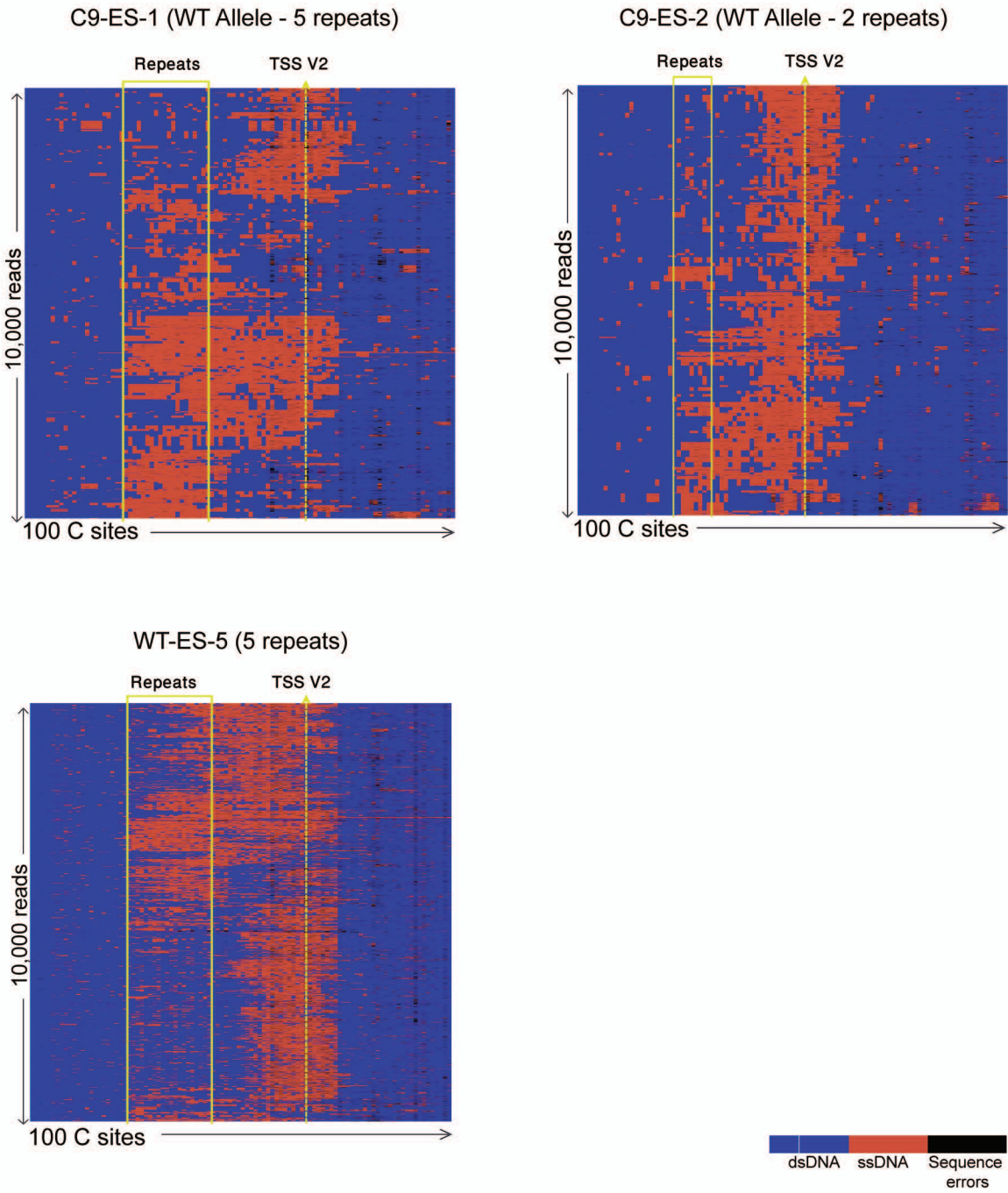

TSS
ATG
FMR1
5'
3'
146,993,370
146,993,551
(CGG)n
146,993,918
146,994,251
60 bp
5' border
3' border
PM-ES-2
uFM-ES-3

Sup 3

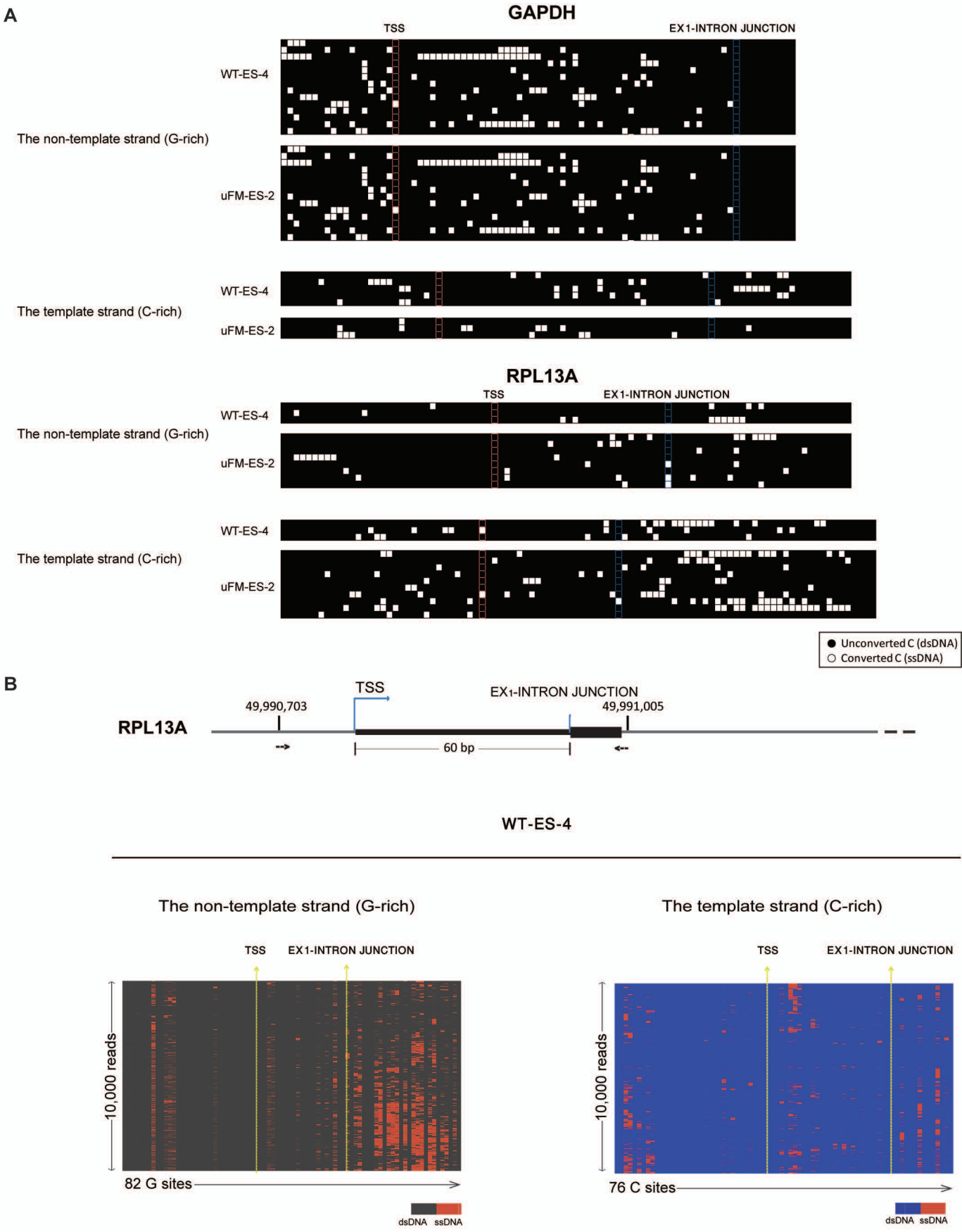

uFM-ES-2
Paternal DNA
216
222
228
234
240
246
1200
800
400
0
223.12
10059
HpaII + HhaI


uFM-ES-2
216
222
228
234
240
246
1200
800
400
0
223.41
9711
229.26
8474
HpaII + HhaI
223.48

Sup 4


C9-ES-1 (WT Allele - 5 repeats)
Repeats
TSS V2
10,000 reads
100 C sites
C9-ES-2 (WT Allele - 2 repeats)
Repeats
TSS V2
10,000 reads
100 C sites
WT-ES-5 (5 repeats)
Repeats
TSS V2
10,000 reads
100 C sites
dsDNA
ssDNA
Sequence errors

Sup 5


A
GAPDH
TSS
EX1-INTRON JUNCTION
WT-ES-4
uFM-ES-2
The non-template strand (G-rich)
The template strand (C-rich)
RPL13A
Unconverted C (dsDNA)
Converted C (ssDNA)
B
RPL13A
49,990,703
TSS
EX1-INTRON JUNCTION
49,991,005
60 bp
WT-ES-4
The non-template strand (G-rich)
The template strand (C-rich)
10,000 reads
82 G sites
76 C sites
dsDNA
ssDNA

SUPPLEMENTARY MATERIAL LEGENDS

Figure S1. Southern blot analysis of unmethylated FX hESC lines. Methylation-sensitive Southern blot analysis of two XX FX hESC lines (u+mFM-ES-4 and uFM-ES-2) during continuous growth in culture. This served to differentiate between unmethylated normal (2.8 kb), premutation (2.9–3.4 kb), and full-mutation alleles (3.4–5.8 kb) and their methylated counterparts, as indicated by 5.2 kb, 5.3-5.8 kb, and fragments larger than 5.8 kb, respectively. Note the tight inverse correlation between methylation and repeat instability in the full mutation range. (B) Relative expression levels of *FMR1* RNA in the hESCs by Taqman qRT-PCR. The expression level in each cell line represents an average of 3-4 independent experiments. Cycle threshold (Ct) values were normalized to the corresponding Ct value of *GUS*. Error bars represent SE (t test for equal variances, $^{*}p<0.05$). For cell lines specifications and expansion size see Table S1.

Figure S2. Mapping single strand DNA displacements on the non-template (G-rich) DNA of *FMR1* by bisulfite DNA footprinting. Bisulfite DNA sequencing data at the boundaries of the single strand DNA displacements in XY hESC with premutation (55<CGGs<200; PM-ES-2) and unmethylated full mutation (CGGs>200 repeats; uFM-ES-3) alleles. Each row represents a single DNA molecule and each circle represents a single cytosine site. Black circles represent bisulfite unconverted Cs (indicating double strand DNA conformation) and white circles represent bisulfite converted C to T (indicating single strand DNA conformation).

Figure S3. Skewed X-inactivation test in uFM-ES-2 hESCs. Skewed X-inactivation (Xi) was confirmed using an established methylation-sensitive quantitative assay, as described in Avitzour et al., 2014. The test is based on digestion with a methylation-sensitive restriction enzyme followed by PCR amplification of a short fragment within the X-linked *ANDROGEN RECEPTOR* gene. The fragment includes a highly polymorphic region (CAG repeat) and several sites that are liable to differential methylation by X-inactivation. Whereas the highly polymorphic CAG repeat is used to distinguish maternal from paternal inherited X-chromosomes, the methylation-sensitive sites allow selective amplification of alleles that are exclusively present on the inactive X chromosome, regardless of parental origin. Accordingly, by comparing the relative amount and fragment size of digested and undigested PCR products using capillary electrophoresis, a skewed bias from the expected 50:50 ratio between the inactive maternal or paternal X chromosomes can be readily identified. Paternal DNA was used to confirm full digestion and to distinguish the maternal (carrying the *FMR1* CGG expansion) from the paternal inherited X chromosome. According to this assay, complete X-inactivation of the maternal X chromosome is evident in uFM-ES-2 hESCs by the detection of a single PCR product of 223bp following digestion with a methylation-sensitive enzyme.

Figure S4. Bisulfite footprinting analysis by deep-sequencing across the *C9orf72* repeats in WT hESCs. DNA bisulfite deep-sequencing across the GGGGCC repeats in *C9orf72* was carried out on native DNA from three different hESC lines (C9-ES-1, C9-ES-2 and WT-ES-5) with a WT allele of 2 or 5 repeats. Bioinformatic analysis resulted in heatmaps representing 10,000 randomly

selected reads, which according to their bisulfite conversion patterns were ascribed to the template strand (C-rich).

Figure S5. Bisulfite footprinting around the TSS in other loci. (A) DNA bisulfite footprinting by single colony sequencing using a pair of unconverted primers resulted in the amplification of the template and non-template DNA strands in two hESC lines (WT-ES-4 and uFM-ES-2) in the region that surrounds the TSS in other, non-repetitive, R-loop forming genes (*GAPDH* and *RPL13A*). TSS and first exon/intron boundaries are marked in red and blue, respectively. (B) Schematic diagram representing the 5'-UTR of *RPL13A* harboring the TSS and the exon/intron junction. DNA bisulfite footprinting by deep-sequencing across the TSS was carried out using unconverted primers in a single hESC line (WT-ES-4). This was followed by a bioinformatic analysis which separated the reads into non-template (G-rich; left) and template strands (C-rich; right), according to conversion patterns.

Table S1

To simplify the labeling of the many different cell lines throughout this manuscript, we designate them as follows:

Cell line	Code name	Original name	Repeat number	REF
XY WT hESCs	WT-ES-1	SZ-13	normal range	(32)
	WT-ES-2	SZ-15		
XX WT hESC	WT-ES-3	HES-123		
XY WT hESC	WT-ES-4	B-200		
XY WT hESC	WT-ES-5	SZ-FX C15B[1]		
hESC with unmethylated full expansion at the FMR1	uFM-ES-1	HEFX	~200-650	
	uFM-ES-2[2]	SZ-FX7	~200-300	
	uFM-ES-3	SZ-FX C12B	~202, 300	
	uFM-ES-4	SZ-FX1		
hESC with premutation expansion at the FMR1	PM-ES-1	SZ-FX4	160[3]	
	PM-ES-2	SZ-FX14 C-11	107, 125[3]	
hESC with an expansion at the C9orf72	C9-ES-1	SZ-ALS1	270	(33)
	C9-ES-2	SZ-ALS3	270	

[1] This cell line was used only as a wild type (WT) control for the *C9orf72* locus

[2] A female cell line with skewed inactivation of the normal allele (32).

[3] Determined using Asuragen (AmplideX® PCR/CE *FMR1*).

[4] Roughly estimated by Southern blot analysis

Table S2 (primers):

Method	Reaction	5' Primer (sequence 5'-3')	3' Primer (sequence 5'-3')	Tm
DRIP	EGR1	GAACGTTCAGCCTCGTTCTC	GGAAGGTGGAAGGAAACACA	60
	RPL13A	AATGTGGCATTTCCTTCTCG	CCAATTCGGCCAAGACTCTA	60
	FMR1	GAGGGCTTCAGGTCTCCTTT	CAGTTGCCATTGTGATTTGG	60
	C9orf72	GCCTCCCCTATTAAGGTTCG	TCTCACGAGGCTAGCGAAAT	60
Colony bisulfite footprinting analysis	FMR1 5' R-loop boundary	GAGGGAACAGCGTTGATCACGTG	TAACAACAACACCTCCATCACC*	56
	FMR1 3' R-loop boundary	GAAGTTTTTTTTTGATTTTGAGAGG*	CCAATGCTAGACCGGAAAAGAG	60
	C9orf72 5' R-loop boundary	AGGAAAGAGAGGTGCGTCAA	CACACAACTCCTAAATTCCAAAAC*	55
	C9orf72 3' R-loop boundary	GTTTTGTTTGGGGGAAGGTT*	GGTGCATGGCAACTGTTTGAATAG	55
	FMR1 repeats sense with R-loops 1 [±]	GAGGGAACAGCGTTGATCACGTG	CCTCTCAAAATCAAAAAAAAACTTCC*	60
	FMR1 repeats sense with R-loops 2	GGAACAGCGTTGATCACGTGACGTGGTTTC		60
	FMR1 repeats (antisense and sense) 1 [±]	GAGGGAACAGCGTTGATCACGTG	CCTCTCGGAGTCGGAGAGGGGCTTC	60
	FMR1 repeats (antisense and sense) 2	GGAACAGCGTTGATCACGTGACGTGGTTTC		61
	C9orf72 repeats (antisense and sense)	TCAGAGAAATGAGAGGGAAAGTAAA	GACCTGATAAAGATTAACCAGAAGAA	58
	RPL13A	GATTGGACATTCGGAAGAGG	GGATGTTAAGCTCCGCAAAA	58
	GAPDH	AAAAGCGGGGAGAAAGTAGG	CTTCAGGCCGTCCCTAGC	58
Next-seq bisulfite footprinting analysis	FMR1 repeats [±]	GAGGGAACAGCGTTGATCACGTG	CCTCTCGGAGTCGGAGAGGGGCTTC	60
		Adapter-GGAACAGCGTTGATCACGTGACGTGGTTTC	Adapter-CACCAGCTCCTCCATCTTCT	61
	C9orf72 repeats [±]	TCAGAGAAATGAGAGGGAAAGTAAA	GACCTGATAAAGATTAACCAGAAGAA	58
		Adapter-TCAGAGAAATGAGAGGGAAAGTAAA	Adapter-GACCTGATAAAGATTAACCAGAAGAA	58
	RPL13A [±]	GATTGGACATTCGGAAGAGG	GGATGTTAAGCTCCGCAAAA	58
		Adapter-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	Adapter-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGA	58
SNP	C9orf72	GTTTTTCCCACCCTCTCTCC	GGTGCATGGCAACTGTTTGAATAG	56

[*] converted primers; [±] hemi-nested reaction