Nucleic Acids Research, 2024, **52**, 4361–4374
https://doi.org/10.1093/nar/gkae124
Advance access publication date: 21 February 2024
Genome integrity, repair and replication

Pathogenic CANVAS $(AAGGG)_n$ repeats stall DNA replication due to the formation of alternative DNA structures

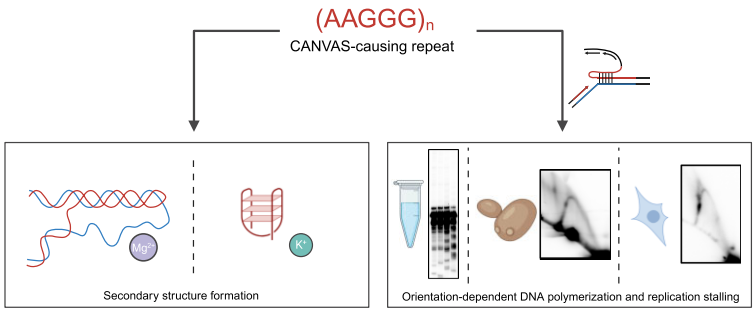
Julia A. Hisey[1], Elina A. Radchenko[1], Nicholas H. Mandel[1], Ryan J. McGinty[2], Gabriel Matos-Rodrigues[3], Anastasia Rastokina[1], Chiara Masnovo[1], Silvia Ceschi[4], Alfredo Hernandez[1], André Nussenzweig[3] and Sergei M. Mirkin[1,*]

[1]Department of Biology, Tufts University, Medford, MA 02155, USA
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[3]Laboratory of Genome Integrity, National Cancer Institute NIH, Bethesda, MD 20892, USA
[4]Department of Pharmaceutical and Pharmacological Sciences, University of Padova, Padova 35131, Italy
[*]To whom correspondence should be addressed. Tel: +1 617 627 4794; Email: sergei.mirkin@tufts.edu

Abstract

CANVAS is a recently characterized repeat expansion disease, most commonly caused by homozygous expansions of an intronic $(A_2G_3)_n$ repeat in the *RFC1* gene. There are a multitude of repeat motifs found in the human population at this locus, some of which are pathogenic and others benign. In this study, we conducted structure-functional analyses of the pathogenic $(A_2G_3)_n$ and nonpathogenic $(A_4G)_n$ repeats. We found that the pathogenic, but not the nonpathogenic, repeat presents a potent, orientation-dependent impediment to DNA polymerization *in vitro*. The pattern of the polymerization blockage is consistent with triplex or quadruplex formation in the presence of magnesium or potassium ions, respectively. Chemical probing of both repeats *in vitro* reveals triplex H-DNA formation by only the pathogenic repeat. Consistently, bioinformatic analysis of S1-END-seq data from human cell lines shows preferential H-DNA formation genome-wide by $(A_2G_3)_n$ motifs over $(A_4G)_n$ motifs. Finally, the pathogenic, but not the nonpathogenic, repeat stalls replication fork progression in yeast and human cells. We hypothesize that the CANVAS-causing $(A_2G_3)_n$ repeat represents a challenge to genome stability by folding into alternative DNA structures that stall DNA replication.

Graphical abstract



Introduction

Biallelic expansions of $(A_2G_3)_n$ repeats cause a newly discovered repeat expansion disease (RED) named CANVAS (cerebellar ataxia, neuropathy, vestibular areflexia syndrome) (1,2). It is an autosomal recessive disease with a carrier frequency range from 0.7% to 4% in studied populations, resulting in a prevalence range from 1:20 000 to 1:625, respectively (1,3). This frequency establishes *RFC1*-related ataxia as likely the most common cause of hereditary late-onset ataxia (1,2,4). CANVAS is a progressive, neurodegenerative disease with a mean age of onset of 52 and a broad spectrum of clinical features, including but not limited to: imbalance, peripheral sensory symptoms, oscillopsia, dry cough, autonomic dysfunction, dysarthria, and dysphagia (4,5).

The expandable $(A_2G_3)_n$ repeat resides in the poly-A tail of an AluSx3 element within the second intron of the *RFC1* gene (1). *RFC1* encodes the largest subunit of replication factor C, the complex responsible for loading PCNA onto DNA during DNA replication and repair (6,7). Given its essential role, it is not surprising that this is the first *RFC1* mutation found to cause human disease. While the pathogenic mechanism of CANVAS remains uncertain, RFC1 loss of function is

Received: July 25, 2023. **Revised:** February 6, 2024. **Editorial Decision:** February 6, 2024. **Accepted:** February 8, 2024

suspected due to (i) its recessive inheritance and (ii) the discovery of CANVAS-affected patients heterozygous for the repeat expansion and an *RFC1* truncating mutation (3,8–11).

Most expandable repeats can adopt alternative, non-B DNA secondary structures that are integral to their propensity to expand and cause disease (reviewed in (12)). These secondary structures are formed during processes involving B-DNA unwinding, such as replication, transcription, and repair, and, at the same time, they can become an obstacle for these processes. Notably, DNA replication and repair have been identified as major sources of repeat instability across various repeats, which has been attributed to their structure-prone nature (reviewed in (12)).

CANVAS differs from most other REDs in that its pathogenic allele varies from the nonpathogenic one not only in repeat size, but also in its base composition. While most healthy individuals harbor $(A_4G)_{11–100}$ repeats in the *RFC1* locus, CANVAS patients largely carry $(A_2G_3)_{250–2000}$ repeats (1,13). Other iterations of the repeat exist, and their pathogenicity is currently being unraveled (1,13–18). Oftentimes, repeat variants are not pure and contain expanded $(A_2G_3)_n$ or other repeat interruptions, adding complexity to assigning the alternative repeat to the pathogenic or benign category (13). Interestingly, the propensity of *RFC1*-containing repeats to expand correlates with the number of guanine residues in the repetitive unit: $A_4G < A_3G_2 < A_2G_3$ (1). Given the repeat composition, we hypothesized the pathogenic repeats can form a stable triplex H-DNA or G-quadruplex DNA, since they are simultaneously homopurine/homopyrimidine (hPu/hPy) mirror repeats (19) and contain evenly spaced G3 runs (20).

Thus, we set out to determine if the pathogenic $(A_2G_3)_n$ repeats form an alternative DNA secondary structure(s), what this structure(s) is, and whether it impedes replication, which *a priori* may lead to the repeat's instability. We found that pathogenic $(A_2G_3)_{10}$ repeats, but not the nonpathogenic $(A_4G)_{10}$ repeats, strongly stall DNA Pol I, Vent, and bacteriophage T7 polymerization *in vitro*. The stalling is orientation-dependent, occurring only when $(A_2G_3)_n$ serves as the template strand. Ambient conditions have a profound effect on the position of the stall, with patterns suggestive of G-quadruplex formation in the presence of potassium or triplex formation in the presence of magnesium. Using chemical probing, we identified the formation of H-r triplex DNA (pyrimidine-purine-purine triplex) (19) by the pathogenic $(A_2G_3)_n$ repeats *in vitro*. Analysis of S1-END-seq peaks at genome-wide H-motifs in human cells revealed that $(A_2G_3)_n$ tracts have a higher propensity than $(A_4G)_n$ tracts to form triplexes. Using two-dimensional electrophoretic analysis of replication intermediates, we show that the pathogenic repeat causes orientation-dependent replication stalling in both yeast and human cells only when the $(A_2G_3)_n$ run is in the lagging strand template. More detailed analysis of replication through $(A_2G_3)_n$ repeats in human cells revealed structures indicative of fork reversal following replication fork stalling. We suggest, therefore, that the non-B DNA-forming potential of the pathogenic repeat during DNA replication results in fork stalling, which may have a role in its instability.

Materials and methods

Specific PCR programs, plasmids, primers and strains used in this study are listed in Supplementary Tables S1–S7.

Plasmid construction

pJH7 was ordered from GenScript and subsequently used as the source of $(A_2G_3)_{60}$ repeats in all downstream cloning. pJH7 consists of a pYES3/CT (Invitrogen) vector with a *URA3* cassette containing an artificial intron with the $(A_2G_3)_{60}$ repeat inserted into the multiple cloning site of the vector. pJH7 is identical to the pYes3-T269-GAA100 vector (21–23) with $(A_2G_3)_{60}$ replacing $(GAA)_{100}$ from this original plasmid.

In vitro primer extension experiments: pJH2 and pJH9 were ordered from GenScript. These plasmids are identical to pJH1 with $(A_2G_3)_{10}$ replacing $(A_2G_3)_{60}$ in pJH1 for pJH2 and $(A_4G)_{10}$ replacing $(A_2G_3)_{60}$ in pJH1 for pJH9. pJH1 plasmid construction is described below in the plasmid construction for yeast two-dimensional gel electrophoresis section.

Plasmid construction for replication analysis in yeast: Primers JH92 and JH93 were used to amplify $(A_2G_3)_{60}$ from pJH7 to add BsrGI restriction sites to the ends of this PCR product and insert this product into plasmid pRS425-UIRLB (24). Clones identified as correct using (i) PCRs flanking the ligation integration site and (ii) Repeat PCR protocol Taq 1 (Supplementary Table S3) were sent for sequencing and the correct clones were named pJH1 (purine lagging strand template) and pJH26 (pyrimidine lagging strand template). Plasmid pJH4 with the $(A_2G_3)_{60}$ repeats replaced by $(A_4G)_{60}$ was ordered from GenScript.

Plasmid construction for replication analysis in human cells: Plasmid pJH5 is identical to the pJC_GAA100 plasmid used for the same purpose previously in our lab (23) with $(A_2G_3)_{60}$ replacing $(GAA)_{100}$. To flip the cassette, we amplified the UR-$(A_2G_3)_{60}$-A3 cassette from pJH5 with primers JH295 and JH296 that incorporate restriction enzyme sites BmgI and BglII on either side of the repeats; these restriction sites were used to re-insert the cassette in the opposite orientation into pJH5. Successfully ligated colonies were screened with PCRs flanking the integration sites. Cloning junctions of correct clones were sequenced, repeat lengths were checked using PCR program Taq 1 (Supplementary Table S3), and the correct plasmid was named pJH6. Gibson assembly with (i) Primers JH53 and JH54 and plasmid pJH4 and (ii) Primers JH55 and JH56 and plasmid pJH5 was used to replace $(A_2G_3)_{60}$ in pJH5 with $(A_4G)_{60}$ from plasmid pJH4. Cloning junctions of correct clones were sequenced, repeat lengths were checked using PCR program Taq 2 (Supplementary Table S3), and the correct plasmid was named pJH11.

Cloning and amplification of $(A_2G_3)_n$ repeats

Due to the secondary structure-forming potential of these repeats, adjustments were made to PCR reaction mixes and cycling programs: Thermo Scientific Phusion High Fidelity DNA Polymerase (Cat# F530L) was used to amplify fragments for Gibson Assembly or cloning. The manufacturer's reaction for Phusion PCR using was altered as follows: the master mix included 1 M Betaine and no DMSO. The program listed in Supplementary Table S1 was followed with the following general changes: the extension time was lengthened to allow for progression through the entire repeat sequence and the extension temperature was raised to 75°C or 80°C when necessary to yield a full-length repeat product. For some PCRs, the annealing temperature was also raised, and longer primers were designed accordingly to ensure annealing. For products with multiple bands used for cloning, the properly sized fragment

was gel extracted and repeat length was confirmed with the Taq repeat PCR programs before cloning. Specific protocols are described in Supplementary Table S2 and generally, annealing and extension temperatures were raised and extension time lengthened. Phusion was also used to amplify the $(A_4G)_n$ repeats. GenScript Taq polymerase (Cat# E00007) was used to amplify the $(A_2G_3)_n$ repeats. The following PCR reaction was used: 1X Taq Buffer with $(NH_4)_2SO_4$ (Thermo Scientific Cat# B33), 1 mM primers, 100 mM dNTP, 1 mM $MgCl_2$, 1 M betaine, 0.625 units Taq polymerase and 1 ng DNA. The PCR programs used for different repeats are listed in Supplementary Table S3.

Supercoiled plasmids used for 2D electrophoresis or chemical probing were run alongside the linearized plasmid on a 0.8% agarose gel with 0.5 μg/ml EtBr (conditions for 7.8 kb plasmid) or 0.6% agarose gel with 1 μg/ml EtBr (conditions for 12.4 kb plasmid) to confirm that their monomeric state. *Escherichia coli* strains with $(A_2G_3)_n$-bearing plasmids were grown at 23°C to reduce instability during plasmid replication.

DNA polymerization

Thermo Sequenase

Double-stranded plasmids pJH2 and pJH9 were used as DNA templates for *in vitro* polymerization experiments with $(A_2G_3)_{10}$ or $(A_4G)_{10}$ repeats, respectively. Primers JH271 and JH272 were used to polymerize through the repeats with either the pyrimidine or purine repeat in the template strand, respectively. Reactions using Thermo Sequenase were carried out as follows: The USBio Thermo Sequenase Cycle Sequencing Kit (Cat# 78500) was used according to the manufacturer's 3′-dNTP internal label cycling sequencing instructions with the following alterations and specifications: Instead of conducting multiple rounds of labeling and extending, 5 μg of each plasmid (1.06 pmol) and 0.5 pmol of primer were used to conduct only one cycle. The primer was pre-annealed to the plasmid in 11 μl water at 95°C for 2 min and immediately submerged in an ice-water bath, then the labeling components were added and the primer extension product was labeled at 60°C for 30 s with 0.5 μl of 1.25 mM [α-^{32}P]dATP. Upon aliquoting this reaction into the four pre-aliquoted termination mixes, termination was carried out at 72°C for 5 min and 4 μl of stop solution was added to each reaction, mixed, and incubated at 95°C for 5 min and immediately submerged into an ice-water bath. Each sequencing termination reaction contained 80 μM dNTPs and 0.8 μM of the designated ddNTP. 8 μl of each reaction was loaded onto a 6% polyacrylamide gel with 7.5 M urea prepared according to manufacturer's instructions (National Diagnostics SEQUAGEL SEQUENCING SYSTEM 2.2 (Cat# EC-833)).

Vent (exo-) polymerase

Reactions using Vent polymerase were carried out as follows: Vent (exo-) DNA polymerase (New England Biolabs Cat# M0257S) was used according to primer extension experiments described in (25) with the following exceptions: 0.5 pmol of primer and 5 μg (1.06 pmol) of DNA were pre-annealed as described above in Thermo Sequenase reactions, labeling was carried out at 60°C for 30 s, termination was carried out at 65 or 80°C for 5 min, and steps after termination are identical to those with Thermo Sequenase. Exact ddNTP and dNTP concentrations varied based on the exact nucleotide to yield optimal sequencing reaction resolution, but ddNTP concentrations ranged from 400 to 900 μM and dNTP concentrations ranged from 30 to 100 μM. The ddNTP:dNTP ratios and concentrations used are described in (26) and in Supplementary Table S4.

T7 DNA polymerase

The protocol found in (27) was followed for the T7 DNA polymerase reactions. In brief, primer/templates used for *in vitro* extension by T7 DNA polymerase were formed by pre-annealing 2.5 μM 5′-^{32}P labeled primer JH270 with 3.125 μM single-stranded oligonucleotides bearing either $(A_2G_3)_{10}$ (JH267), $(T_2C_3)_{10}$ (JH268), or $(A_4G)_{10}$ (JH269) repeats in 10 mM Tris–HCl, pH 7.5, 0.1 mM EDTA alone or with the addition of various salts (50 mM KCl, LiCl, and NaCl) by incubating at 95°C for 5 min, followed by a gradual return to room temperature. Primer extension reactions consisted of 10 nM primer/template and 1 μM T7 DNA polymerase in 40 mM Tris–HCl, pH 7.5, 5 mM dithiothreitol (DTT), 0.3 mM dNTPs, and 55 mM of the same monovalent metal chloride salt used in the annealing buffer (LiCl, KCl, or NaCl). DNA polymerization was initiated by addition of 10 mM $MgCl_2$ for all reactions, followed by incubation at room temperature for 1 min, and reactions were quenched with formamide/EDTA loading dye. Either a ladder or a sequencing reaction was used to determine the location of stalls within the repeats. The ladder was made by 5′-end labeling single-stranded oligos containing 2, 4, 6, 8 or 10 $(C_3T_2)_n$ repeats (oligos found in Supplementary Table S5), which are the exact products that would result from primer extension reactions terminating on an $(A_2G_3)_{10}$ template at these different repeat units. The sequencing reactions were carried out using Thermo Sequenase as in the *in vitro* Thermo Sequenase protocol described above with the single-stranded oligos and primers for the *in vitro* T7 experiments. Primer extension products, ladders, and sequencing reactions were resolved by electrophoresis on a denaturing 10% polyacrylamide gel.

Chemical probing

2 μg of the repeat-containing pJH2 or pJH9 plasmid (0.424 pmol) was incubated in a 10 mM Tris–HCl pH 7.5 2 mM $MgCl_2$ buffer with either 6 mM $KMnO_4$ or the same volume of water for 2 min at 37°C. The reaction was quenched with 1 M β-mercaptoethanol, precipitated with 100% ethanol, rinsed with 70% ethanol, and resuspended in 5 μl water. The primer extension was carried out using the USBio Thermo Sequenase Cycle Sequencing Kit (Cat# 78500) protocol with the same changes as was described in the *in vitro* Thermo Sequenase polymerization methods with the following alterations: the chemically probed DNA was pre-annealed with 0.5 pmol of primer JH271 for a final volume of 5.5 μl, the labeling reaction's final volume was 8.75 μl with the same ratios as described in the manufacturer's instructions, and dNTPs were added to a final concentration of 75 μM for each dNTP for extension.

Data analysis of S1-END-seq, P1-END-seq and BG4-ChIP-seq

S1-END-seq and P1-END-seq sequencing reads were aligned as previously described (28). In brief, reads were aligned to the reference genome (hg19) using bowtie (v.1.1.2) (29) with parameters -n 3 -l 50 -k 1. Samtools (v.1.11) (30) were

used to convert and sort the aligned.sam files to sorted.bam files..bam files were further converted to bed files using bedtools (v.2.29.2) (31). The list of $(A_2G_3)_n$ repeats annotated in the hg19 human genome is provided in Supplementary Table S8.

MACS 1.4.3 (32) was used with the parameters -p 1e-5 –nolambda—nomodel –keep-dup = all to call peaks and the peaks were then filtered by a 10-fold enrichment over background.

For Figure 3A and Supplementary Figure S3: Shared S1-END-seq peaks from five cell lines were obtained from a previous study (28) and provided as BED files in Supplementary Files S1-S5, which represent the output of the 'Peak calling' stage. Custom Python scripts were written to perform the following analysis. Overlapping genomic coordinates from the five cell lines were combined into non-overlapping peaks. Coordinates were converted to GRCh38 using Liftover. A set of control coordinates were randomly generated, matching the S1-END-seq peaks in total number, the length of each peak, and the proportion located on each chromosome. S1-END-seq peaks and control coordinates were compared to a database of repetitive sequences generated previously (33). For each category or subtype of repeat, the proportion of those falling within 100 nucleotides of the S1-END-seq peaks or control coordinates was calculated for each repeat length. Linear best fit lines for each distribution were weighted to the total number of repeats in each length bin and were also restricted to bins containing at least three repeats. Comparison of distributions along repeat lengths were made by Wilcoxon signed-rank test, restricted to length bins containing at least three repeats, $P < 2.7 \times 10^{-7}$. Because these data are paired non-normal distributions, the non-parametric Wilcoxon test performs an appropriate statistical comparison.

BG4-ChIP-seq peaks from HACAT cells (34) were aligned to the hg19 reference genome using Bowtie (version 1.1.2) (29) with the options-best-all -strata-l 50 allowing 2 mismatches and discarded tags with multiple alignments (-n 2 -m 1). Peak calling was performed using the previously described settings in (34) (https://github.com/sblab-bioinformatics/dna-secondary-struct-chrom-lands/blob/master/Methods.md#g4-chip-peak-calling).

Reference datasets

S1_END_seq in KM12 paired with P1-END-seq (GSM6372152), P1_END_seq_KM12 (GSM6372153), S1_END_seq in lymphoblast GM15851 paired with P1_END_Seq (GSM6181010), P1_END_seq in lymphoblast GM15851 (GSM6181011). Asynchronous KM12 cells (GSM6180996) were compared with experimentally paired G1 arrested KM12 cells (GSM6181003). S1-END-seq performed in HACAT cells (GSM6195485). BG4-ChIP-seq (GSM2035782) and input-BG4-ChIP-seq (GSM2035782).

Strains for the analysis of replication in yeast

JAH231 (*MATa leu2-Δ1, trp1-Δ63, ura3–52, his3–200, bar1Δ*) is the background strain used for plasmid replication for all replication analysis in yeast. 3 *pif1Δ* strains were constructed in the JAH231 background using simple gene replacement with the hygromycin resistance gene using the primers described in Supplementary Table S6. Primers listed in Supplementary Table S6 were used to check the 5′ and 3′ flanks and for absence of the *PIF1* gene. Plasmids described above that were constructed for this purpose contained the repeats, a yeast 2μ origin of replication, and the ampicillin resistance gene. Plasmids were named pJH1, pJH26, and pJH4 and have $(A_2G_3)_{60}$, $(C_3T_2)_{60}$ or $(A_4G)_{60}$ in the lagging strand template in relation to the yeast 2μ origin. After plasmids were transformed into a given strain, the repeat size was checked.

Analysis of replication intermediates by two-dimensional (2D) gel electrophoresis in yeast and human cells

Yeast cells: Following the methods outlined in (24), strains containing the yeast two-dimensional gel plasmid were grown, yeast replication intermediates were extracted, digested with restriction enzymes, run on 2D gel and analyzed.

Human cells: The plasmids used for human cell two-dimensional gel electrophoresis contain the repeats, the SV40 origin of replication, the gene encoding the large T antigen, and the ampicillin resistance gene. Plasmids were named pJH5, pJH6, and pJH11 and have $(A_2G_3)_{60}$, $(C_3T_2)_{60}$, or $(A_4G)_{60}$ in the lagging strand template in relation to the SV40 origin of replication and their construction is more fully described above. Following the methods outlined in (23), plasmids were transfected into HEK293T cells, incubated for 48 hours, replication intermediates were extracted, digested with restriction enzymes AfiII and KpnI, and intermediates were run on a 2D gel. For a more detailed analysis of replication intermediates, cells 48 hrs post-transfection were spun down and resuspended in 1.5 ml Hirt lysis buffer (10 mM Tris–HCl pH 7.5, 10 mM EDTA pH 8.0, 0.6% SDS). The mixture was treated with 200 μg of proteinase K at 37°C for 90 min. NaCl was then added to a final concentration of 1 M and the mixture was incubated overnight at 4°C. The following day, the mixture was spun down and the supernatant was further purified using phenol:chloroform:isoamyl alcohol followed by isopropanol precipitation. Replication intermediates were digested with appropriate restriction enzymes for 9 h at 37°C as to place repeats on descending arm of Y-arc. Samples of lagging strand synthesis through $(A_2G_3)_{60}$ were digested with DpnI, XbaI and BsrGI (New England Biolabs) to yield a 2.6 kb fragment, while samples of lagging strand synthesis through $(C_3T_2)_{60}$ were digested with DpnI, PciI, and BsrGI to yield a 2.9 kb fragment.

Two-dimensional gel electrophoresis quantification was carried out as described in (24) and demonstrated visually in Supplementary Figure S7. For each biological replicate, of which there were three for yeast and three for human cell replication, the following was performed. Using ImageJ, the Y-arc was traced from the 1.5n spot to the 2n spot when the repeats are on the descending arm of the Y-arc and from the 1n spot to the 1.5n spot when the repeats are on the ascending arm of the Y-arc in order to have a measure of the density of the arc. The same was done for a line above and below the arc to have an average background signal. The distance and density data were graphed using RStudio to produce a graph such as the example in Supplementary Figure S7 using the code referenced in (24). Using ImageJ, lines were drawn to estimate where a smooth arc profile would be (35) and between this smooth arc and the averaged background line, which was calculated using RStudio. The area between the stall and the estimated smooth arc line (labeled 1 in Supplementary Figure S7) and the area between the estimated smooth arc line and the background line (labeled 2 in Supplementary Figure S7) are used to

quantify the stall with the formula, area 1/(area 1 + area 2), which is the number seen in the bar graphs. For all two-dimensional gel electrophoresis quantification, at least three independent experiments were conducted for each orientation or repeat. Statistical analysis was performed using GraphPad Prism.

Statistical analysis and data visualization

For other data visualization, we used GraphPad prism or R studio. Statistical analyses are specified in the figure legends.

Results

Pathogenic $(A_2G_3)_n$ repeats strongly stall DNA polymerization *in vitro* in an orientation-dependent manner

Structure-forming repeats pose a significant obstacle to polymerases for DNA synthesis during DNA replication and repair (36–42). This phenomenon is believed to be a driver of repeat instability (reviewed in (12,43,44)). Replication issues are more severe for many repeats when the structure-prone strand is on the lagging strand template, which is thought to be due to the single-stranded Okazaki initiation zone, allowing for structure formation. In line with this reasoning, repeats are often particularly unstable in this orientation (reviewed in (12)). To the best of our knowledge, no experimental data are available on the replication nor instability of the expandable CANVAS $(A_2G_3)_n$ repeat.

We, thus, studied DNA polymerization with both A- and B-family polymerases through the pathogenic or nonpathogenic repeats in both orientations. First, we used Thermo Sequenase, a mutated DNA Polymerase I (exo-) (45), and a repeat-containing double-stranded plasmid template with primers that anneal up- or downstream of the repeats. The DNA sequencing reactions were carried out at 72°C in the presence of 3.5 mM Mg^{2+} as described in Materials and methods. Note that upon denaturing and rapid primer annealing, the plasmid template becomes a coil of interlinked single-stranded DNA segments, rather than a plain circular double-stranded DNA (Figure 1C). When the $(A_2G_3)_{10}$ run serves as the template, polymerization stalls profoundly at its center (at the fifth and sixth repeats) and is unable to progress further in almost all templates (Figure 1A, panel 1). In contrast, when the $(C_3T_2)_{10}$ repeat serves as the template, polymerization progresses through the repeats smoothly (Figure 1A, panel 2). For the nonpathogenic repeats, when the $(A_4G)_{10}$ run is in the template strand, polymerization only mildly stalls at the sixth and seventh repeats, while the majority of DNA polymerases progress through the repeats (Figure 1A, panel 3). The pyrimidine $(CT_4)_n$ run in the template strand does not pose an obstacle for the DNA polymerase (Figure 1A, panel 4).

A priori such a structure could either be (i) an H-r DNA triplex formed when DNA polymerase reaches the center of the template, or (ii) a G-quadruplex formed by the G-rich template strand. The observed DNA polymerase stalling at the center of the $(A_2G_3)_{10}$ strand strongly implicates H-r DNA triplex formation (Figure 1D). The lack of a potassium cation due to optimal Thermo Sequenase reaction conditions renders G-quadruplex formation less likely (46,47). Furthermore, G-quadruplex-forming sequences usually stall polymerases directly at the 3′ end of the sequence (47,48), rather than at the center. The minor stall within the nonpathogenic $(A_4G)_{10}$ repeat template is likely caused by a much weaker H-r DNA triplex. It was indeed found that A-rich H-r triplexes are stabilized by Zn^{2+}, rather than Mg^{2+} cations (49).

We next asked if these results would translate to B-family DNA polymerases, as this family also encompasses eukaryotic replicative polymerases. To address this question, we conducted DNA sequencing with Vent (exo-) DNA polymerase using the same templates and primers that were used with Thermo Sequenase. We observed that Vent polymerase stalls profoundly during polymerization through the pathogenic $(A_2G_3)_{10}$ repeats in a magnesium-, orientation-, and temperature-dependent manner (Figure 1B, Supplementary Figure S1). Vent progresses smoothly through the repeats with $(A_2G_3)_{10}$ as the template at 80°C with 2 mM Mg^{2+} (Figure 1B, panel 1). When the magnesium concentration is increased to 5 mM, Vent stalls in the middle of the repeats, indicative of a triplex-mediated stall as seen with Thermo Sequenase (Figure 1B, panel 2). Notably, Vent does not stall as it progresses through the nonpathogenic repeats with $(A_4G)_{10}$ in the template strand at these conditions (Figure 1B, panel 5). Similarly, Vent does not stall significantly as it progresses through the pathogenic repeats with $(C_3T_2)_{10}$ serving as the template even at 65°C (Figure 1B, panel 4). In contrast to the Thermo Sequenase buffer, the Vent extension buffer contains 10 mM KCl, allowing for the possibility of G-quadruplex formation as well. Strikingly, when the temperature is lowered from 80°C to 65°C with either 2 or 5 mM Mg^{2+}, Vent stalls at the beginning of the repeats (Figure 1B, panel 3, Supplementary Figure S1). It is likely a G-quadruplex-mediated stall, given a combination of its magnesium-independence and location at the beginning of the repeats (Figure 1E). When the magnesium concentration is raised from 2 to 5 mM at 65°C, DNA sequencing reactions show stalling both at the beginning and in the middle of the repeat (Figure 1B, panel 3, Supplementary Figure S1). Therefore, at low temperatures and in the presence of both potassium and magnesium, Vent is stalled due to either triplex or G-quadruplex formation (Figure 1D, E).

We then investigated DNA polymerization through single-stranded CANVAS repeat-containing templates by bacteriophage T7 DNA polymerase. T7 DNA polymerase is a robust mesophilic enzyme that is often used as an *in vitro* model system for processive DNA polymerases and replication forks (50). Differently from Thermo Sequenase and Vent polymerase, it is active at 25°C in the presence of both potassium and magnesium ions. For this initial experiment, we annealed a primer to a single-stranded template upstream of 10 repeat units: either $(A_2G_3)_{10}$, $(C_3T_2)_{10}$ or $(A_4G)_{10}$, followed by DNA polymerization. T7 DNA polymerase only stalls when $(A_2G_3)_{10}$ serves as the template strand and does not stall when $(C_3T_2)_{10}$ or the nonpathogenic $(A_4G)_{10}$ repeat serve as the template strand (Supplementary Figure S2A). Notably, in the presence of 10 mM Mg^{2+} and in the absence of potassium in the extension reaction, T7 stalls about halfway through the repeats (around the sixth repeat) (Supplementary Figure S2A). When the magnesium concentration is kept constant, but 55 mM K^+ is added to the extension reaction, the stalling occurs at the beginning of the repeats (Supplementary Figure S2A). Given the ion-dependence and earlier-described reasoning regarding the position of the stall, we believe this is indica-

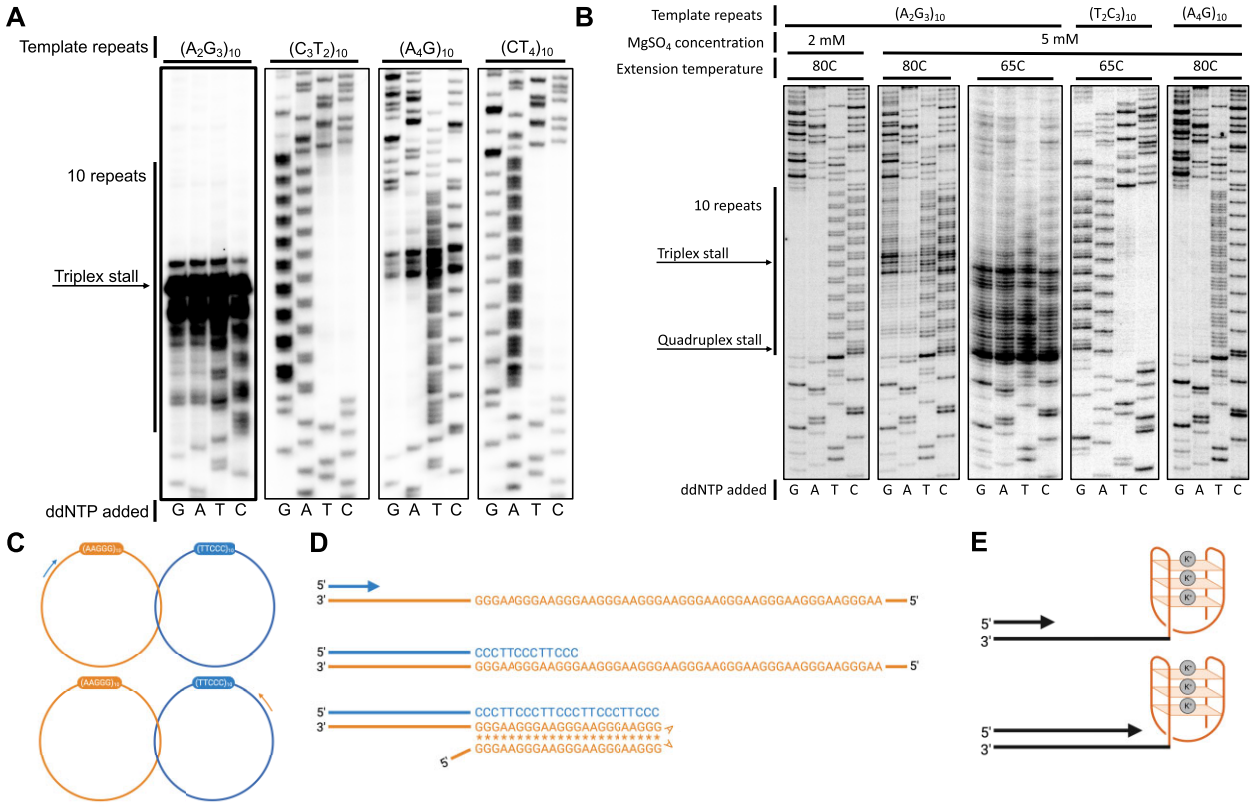


Figure 1. Thermo Sequenase and Vent (exo-) polymerases stall during polymerization through the pathogenic $(A_2G_3)_{10}$ repeat *in vitro*. (**A**) Sequencing reaction by Thermo Sequenase stalls in the middle of the pathogenic $(A_2G_3)_{10}$ repeats when they serve as the template strand. Polyacrylamide gel electrophoresis separation of Thermo Sequenase sequencing reactions performed as described in Materials and Methods. Briefly, 5 μg of each plasmid and 0.5 pmol of primer were denatured and pre-annealed and the USBio Thermo Sequenase Cycle Sequencing Kit's 3′-dNTP internal label cycling sequencing instructions were followed. Radioactive labeling was performed at 60°C for 30 s and primer extension reactions were performed at 72°C for 5 min. (**B**) Sequencing reaction by Vent polymerase stalls at the beginning or middle of the pathogenic $(A_2G_3)_{10}$ repeats, depending on surrounding ions, when they serve as the template strand. Polyacrylamide gel electrophoresis separation of Vent sequencing reactions were performed as described in Materials and Methods. Briefly, 5 μg of each plasmid and 0.5 pmol of primer were denatured and pre-annealed, radioactive labeling was carried out at 60°C for 30 s, and primer extension was carried out at 65 or 80°C for 5 min. (**C**) Schematic of denatured, intertwined double-stranded plasmid with primers annealed to allow for primer extension reactions through the purine- or pyrimidine-rich strand of $(A_2G_3)_{10}$ or $(A_4G)_{10}$ in the template strand. Primers are 98 base pairs or 75 base pairs away from repeats for the purine-rich or pyrimidine-rich template, respectively. Created with BioRender. (**D**) Model for triplex formation as polymerase progresses through the repeats with the purine-rich strand as the template. Created with BioRender. (**E**) Model for G-quadruplex formation as the polymerase reaches the beginning of the repeats with the purine-rich strand as the template. Created with BioRender.

tive of a triplex-mediated stall without potassium (Figure 1D) and G-quadruplex-mediated stall in the presence of potassium (Figure 1E).

Altering the annealing and primer extension buffer conditions to include various monovalent ions revealed that the T7 polymerase stalling pattern with $(A_2G_3)_{10}$ in the template strand changes depending on the surrounding ion concentration (Supplementary Figure S2B). When lithium or sodium is added instead of potassium, T7 polymerase progresses smoothly through the repeats without stalling (Supplementary Figure S2B). These monovalent ions are known to not stabilize G-quadruplexes as potassium does, thereby supporting the conclusion that the stall at the beginning of the repeats is G-quadruplex-mediated. No stalling is observed when $(C_3T_2)_{10}$ or $(A_4G)_{10}$ is in the template strand, regardless of surrounding ions (Supplementary Figure S2C).

Altogether, we conclude that depending on ambient conditions, either an H-r triplex or G-quadruplex is formed during polymerization of the $(A_2G_3)_{10}$ template, which blocks DNA polymerase progression *in vitro*.

$(A_2G_3)_n$ repeats form triplex DNA in supercoiled DNA

While both the pathogenic $(A_2G_3)_n$ and nonpathogenic $(A_4G)_n$ repeats are hPu/hPy mirror repeats, the former is G-rich and the latter is A-rich. *A priori*, both can form triplex H-DNA, but the $(A_2G_3)_n$ repeat would most likely form an H-r triplex structure at physiological pH given the cytosine protonation required for the H-y isoform, while the $(A_4G)_n$ repeat could form either the H-r or H-y isoform, if it is able to form a triplex (19,51). In addition, the $(A_2G_3)_n$ repeat has the potential to form G4-DNA as it has regularly spaced G3 blocks. The $(A_4G)_n$ repeat, in contrast, can convert into the so-called propeller DNA (P-DNA) (52) or be a DNA unwinding element (DUE) as it has multiple A_n-runs (52,53).

To distinguish which of these structures are formed by those repeats in supercoiled DNA, we mapped single-stranded portions within the repeats *in vitro*. We used potassium permanganate ($KMnO_4$), which preferentially modifies single-stranded thymines (54), thereby blocking Watson–Crick hydrogen bonding with a complementary strand and allowing

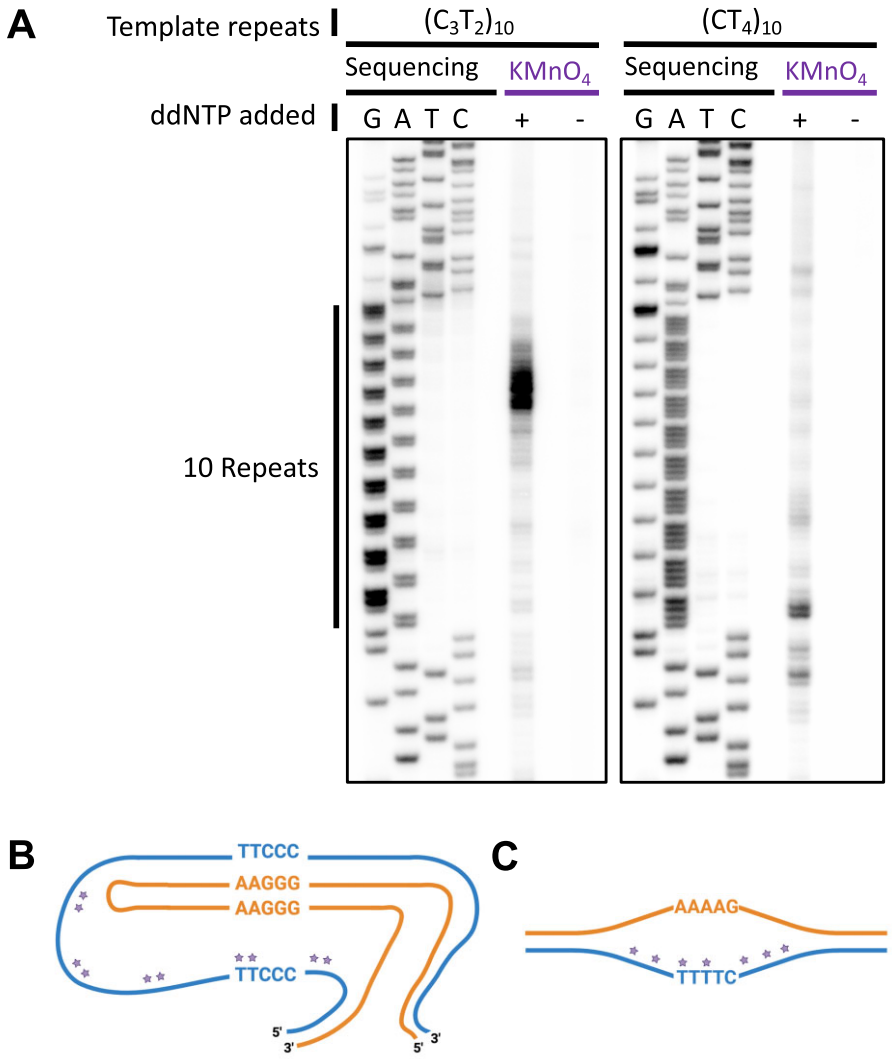


Figure 2. Potassium permanganate probing of pathogenic and nonpathogenic repeats reveals H-r3 formation by the $(A_2G_3)_{10}$ repeat. (**A**) Polyacrylamide gel electrophoresis separation of sequencing reactions and primer extension reactions on potassium permanganate- or water-treated repeat-containing plasmids using the pyrimidine-rich strand as a template. 5 μg of repeat-containing supercoiled DNA was incubated in 10 mM Tris–HCl pH 7.5, 2 mM $MgCl_2$ buffer with either 6 mM $KMnO_4$ or the same volume of water for 2 min at 37°C. The reaction was quenched, resuspended in water, and used as a template for primer extension as described in the Materials and Methods. The Thermo Sequenase sequencing reactions alongside were performed as described in Figure 1. (**B**) H-r3 triplex predicted from chemical probing for $(A_2G_3)_{10}$ repeats. (**C**) DNA unwinding element (DUE) predicted from chemical probing for $(A_4G)_{10}$ repeats. Purple stars in (B) and (C) represent possible $KMnO_4$ modification sites. Created with BioRender.

their detection as polymerization termination sites in a primer extension assay. This approach could not be used for the $(A_2G_3)_n$ strand since it stalls the polymerase, but interrogating the $(C_3T_2)_n$ strand would allow us to distinguish between the candidate structures.

Upon potassium permanganate modification of plasmids with $(A_2G_3)_{10}$ repeats, Thermo Sequenase terminates in the wide area between the sixth and ninth repeats (Figure 2A). Meanwhile, the unmodified plasmid shows an almost undetectable level of polymerase termination (Figure 2A). This data shows that the 5′-half of the polypyrimidine strand of the pathogenic repeat is single-stranded, consistent with the H-r3 DNA triplex (19) (Figure 2B). At the same time, $KMnO_4$ modification of the $(A_4G)_{10}$ repeat in supercoiled DNA shows evident termination signals at the beginning of the repeat combined with weak termination signals throughout the repeat (Figure 2A). This pattern is a stark contrast from the pathogenic repeat and suggests the $(A_4G)_{10}$ repeats in supercoiled DNA may be transiently unwound, which is consistent with DUE chemical modification (53) (Figure 2C). Note, however, that the buffer used in the chemical probing experiments had a very low ionic strength, which may favor unwinding of the AT-rich $(A_4G)_{10}$ repeat over triplex formation. It is possible that the $(A_4G)_n$ repeat could still form a triplex at higher ionic strengths. Indeed, the weak stall observed at $(A_4G)_{10}$ repeats during Thermo Sequenase polymerization *in vitro* indicates the $(A_4G)_n$ repeat has the ability to form a triplex, albeit a weaker one than the $(A_2G_3)_n$ repeat.

$(A_2G_3)_n$ repeats have a higher propensity to form triplex DNA *in vivo* than $(A_4G)_n$ repeats

To examine the triplex-forming potential of $(A_2G_3)_n$ and $(A_4G)_n$ repeats genome-wide in human cell lines, we reanalyzed an S1-END-seq dataset from a previous study (28). This method identifies single-stranded DNA regions in the

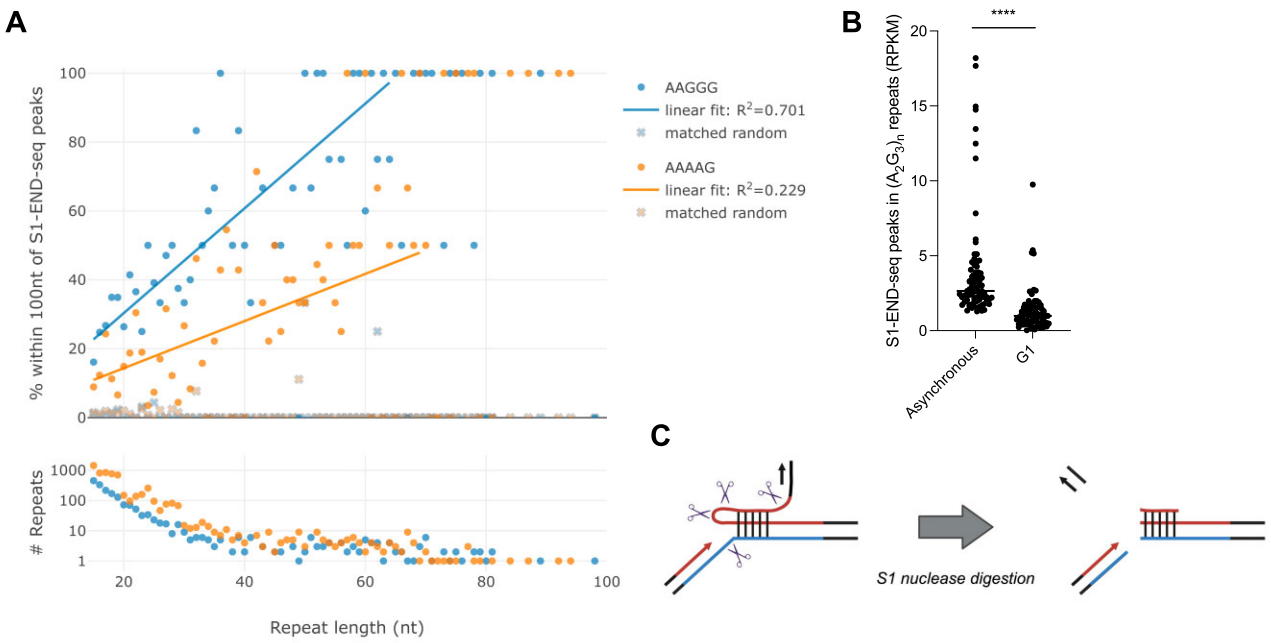


Figure 3. Triplex-formation in $(A_2G_3)_n$ repeats genome-wide in human cells. S1-END-seq data from Matos-Rodrigues *et al.* 2022 (28) were used for these analyses. (**A**) Comparison of pathogenic $(A_2G_3)_n$ and $(A_4G)_n$ motifs genome-wide. Overlapping S1-END-seq peak genomic coordinates from the five cell lines were combined into non-overlapping peaks and the coordinates converted to GRCh38. A set of control coordinates were randomly generated, matching the number of S1-END-seq peaks in total number, length of each peak, and proportion on each chromosome. S1-END-seq peaks and control coordinates were compared to a previously generated repetitive sequence database (33). For each category or subtype of repeat, the proportion of those falling within 100 nucleotides of the S1-END-seq peaks or control coordinates was calculated for each repeat length and graphed on the Y-axis. Best-fit line and R^2 value produced using R. Top: graph depicting the percentage of repeats found within 100 nucleotides of S1-END-seq peaks as the repeat length increases. Bottom: graph depicting the number of repeat sequences of each repeat length as the repeat length increases. The X-axis of the upper and lower panel of each graph is the repeat length. Comparison of distributions along repeat lengths were made by Wilcoxon signed-rank test, restricted to length bins containing at least three repeats, $P < 2.7 \times 10^{-7}$. The test was applied via the Scipy package in Python (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html#scipy.stats.wilcoxon). (**B**) S1-END-seq peaks from Matos-Rodrigues *et al.* (28) that overlap with $(A_2G_3)_n$ repeats annotated in hg19 human genome in asynchronous or G1-arrested (via CDK4/6 inhibitor Palbociclib 10mM for 24 h) KM12 cells. Data and experimental methods can be found in Matos-Rodrigues *et al.* (28). RPKM = reads per kilobase per million mapped reads. KM12 cells were compared using the Mann–Whitney U test, **** $P < 0.0001$. (**C**) Representation of S1 nuclease cleavage of triplex H-DNA, yielding one double-stranded end available for sequencing adapter ligation and one triple-stranded end that is unavailable for sequencing adaptor ligation. The red line indicates the homopurine strand, while the blue line indicates the homopyrimidine strand of an H-r triplex, although the same end products would exist for an H-y triplex.

human genome, commonly resulting from the formation of $(AT)_n$ DNA cruciforms and H-DNA triplexes (28). The S1-END-seq technique uses S1 nuclease to convert single-stranded DNA to double-strand breaks, which are then used as substrates for the attachment of high-throughput sequencing adapters (28). Cleavage of H-DNA triplex structures yields one double-stranded and one triple-stranded end, thereby only producing one end (the double-stranded end) that a sequencing adaptor can be ligated to (illustrated in (28) and Figure 3C). Therefore, the S1-END-seq pattern is asymmetric for H-DNA triplexes as opposed to other alternative DNA structures.

To evaluate the triplex-forming potential of $(A_2G_3)_n$ and $(A_4G)_n$ repeats, we compared the genomic coordinates of all such repeats in the human genome to the coordinates of peak regions identified by S1-END-seq. For both $(A_2G_3)_n$ and $(A_4G)_n$ repeats, we see frequent overlap with the S1-END-seq peaks, growing to as much as 90% for very long repeat tracts, while overlap with randomly-generated genomic coordinates is in line with the 1.3% of the genome contained within peaks (Figure 3A). In contrast, G4-DNA motifs are not highly enriched in S1-END-seq peaks, with only 2–3% appearing in peaks regardless of motif length (Supplementary Figure S3B). Other non-B-forming motifs are similarly not enriched in the S1-END-seq peaks (Supplementary Figure S3C), with the sole exception of $(AT)_n$ repeats (Supplementary Figure S3D), demonstrating that the S1-END-seq assay is primarily specific for triplex-forming DNA at H-DNA motifs, especially since these motifs are unable to form the cruciform structures identified at $(AT)_n$ repeats. Thus, it is highly likely that both $(A_2G_3)_n$ and $(A_4G)_n$ repeats form DNA triplexes genome-wide in human cell lines. We then asked whether $(A_2G_3)_n$ repeats would overlap with sites of G4-DNA formation. Therefore, we re-analyzed BG4-ChIP-seq data from Hänsel-Hertsch *et al.* 2016 performed in the HACAT cell line (34). None of the 21 278 peaks found in BG4-ChIP-seq overlapped with $(A_2G_3)_n$ repeats annotated in the hg19 human genome (Supplementary Figure S4).

Strikingly, $(A_2G_3)_n$ repeats demonstrate consistently higher triplex-forming potential than $(A_4G)_n$ repeats along the axis of repeat length (Wilcoxon one-sided test, $P < 2.7 \times 10^{-7}$) (Figure 3A). Furthermore, other repeats known to be pathogenic—$(AG_4)_n$ and $(A_3G_2)_n$—also demonstrate higher triplex-forming potential than $(A_4G)_n$ repeats, all displaying a similar trend as $(A_2G_3)_n$ (Supplementary Figure S3A). Looking at a variety of hPu/hPy motifs with increasing numbers

of adenines in a row, we see a pattern emerging, in which four or more adenines in a row becomes detrimental to triplex stability (Supplementary Figure S3E). Looking at hexanucleotide motifs, though power-limited, we see that $(A_4G_2)_n$ and $(A_5G_1)_n$ show few signs of triplex formation, unlike all other hPu/hPy hexanucleotides (Supplementary Figure S3F). Though $(A_4G_2)_n$, $(A_3GAG)_n$ and $(A_2G)_n$ motifs all contain the same A:G ratio, only $(A_4G_2)_n$ motifs are inhibited in triplex formation (Supplementary Figure S3G). We hypothesize that the presence of longer adenine runs in double-stranded DNA favors stiff propeller DNA (P-DNA) (52), making triplex nucleation problematic. $(A)_n$ repeats clearly do not form triplexes, while $(G)_n$ repeats show a slight elevation in overlaps with S1-END-seq peaks (Supplementary Figure S3H), consistent with our observations at other G4-DNA motifs (Supplementary Figure S3B).

The S1-END-seq protocol requires acidic conditions, which are known to stabilize YRY (pyrimidine-purine-pyrimidine) triplexes, potentially confounding our conclusions (28). Matos-Rodrigues *et al*. have used several experimental assays to demonstrate that S1-END-seq reveals H-DNA formed *in vivo* and not during sample processing (28). One of these assays was a new method, P1-END-seq, wherein P1 nuclease is used instead of S1 nuclease and therefore single-stranded DNA cleavage can occur at a neutral pH. We overlapped the S1-END-seq and P1-END-seq peaks that occur at $(A_2G_3)_n$ repeats annotated in the hg19 human genome and found that over 80% of the P1-END-seq peaks overlapped with the S1-END-seq peaks in the two cell lines tested (Supplementary Figure S5). Note that the smaller number of triplexes detected by P1-END-seq compared to S1-END-seq may reflect H-y triplex formation during sample preparation for S1-END-seq, or it could be attributed to the fact that P1-END-seq is a newer and less-optimized method.

We further reasoned that if S1-END-seq was identifying H-y DNA triplexes formed by $(A_2G_3)_n$ repeats during sample preparation, the signal should be similar at different stages of the cell cycle. However, our experimental data shows exactly the opposite: the profile of S1-END-seq peaks at $(A_2G_3)_n$ repeats genome-wide differs dramatically between asynchronous and G1-arrested KM12 cells (Figure 3B). The S1-END-seq signal at $(A_2G_3)_n$ repeats is almost abolished upon G1 arrest (Mann–Whitney *U* test, $P < 0.0001$) (Figure 3B), demonstrating that $(A_2G_3)_n$ repeats preferably adopt triplex conformation *in cellulo* in cycling cells. Notably, if a triplex is formed during DNA replication, it would also yield one double-stranded and one triple-stranded end upon S1 nuclease digestion, resulting in an asymmetric S1-END-seq signal (Figure 3C). The stability of the triple-stranded end is likely due to slow dissociation kinetics of triplexes (particularly long purine-purine-pyrimidine triplexes) and to the degradation of a DNA strand complementary to the third strand of the triplex. In contrast, S1 nuclease cleavage of single-stranded gaps formed during DNA replication through a repeat in duplex form would result in two double-stranded ends (Supplementary Figure S6).

Pathogenic $(A_2G_3)_n$ repeats stall replication in yeast in an orientation-dependent manner

In vitro polymerization through the repeats suggests stable H-r triplex or G-quadruplex formation as the polymerase progresses through the pathogenic repeats. However, it is not clear which structure would prevail during DNA replication in cells, given two major factors: (i) far more components are at play in the replication fork as compared to DNA polymerase alone and (ii) intranuclear conditions, including chromatin, DNA supercoiling, and ion concentrations, differ from that in the polymerization reactions. We hypothesized, based on our chemical probing and bioinformatic analysis, that the pathogenic $(A_2G_3)_n$ repeats form H-r triplex DNA in cells and would therefore stall replication only when the purine-rich strand resides on the lagging strand template. We also hypothesized that the nonpathogenic $(A_4G)_n$ repeats would not stall the replication fork to the same extent as the pathogenic repeats.

To study whether the expandable CANVAS repeats stall DNA replication in yeast, we cloned a longer $(A_2G_3)_{60}$ repeat tract into the multicopy yeast pRS425 plasmid in both orientations with respect to the replication direction and analyzed their replication using two-dimensional electrophoretic analysis of replication intermediates (Figure 4) as described in Krasilnikova *et al*. (35). We chose to use a multicopy plasmid as this would yield more replication intermediates. We believe that this approach is justified since our previous electrophoretic analyses of replication intermediates did not show any difference in the pattern of fork stalling at $(GAA)_n$ repeats whether they were located on a multicopy plasmid or within chromosome III (42,55).

There is no observable fork stall in the no repeat control plasmid (Figure 4A, B). At the same time, the pathogenic repeat stalls replication fork progression only when its homopurine run is in the lagging strand template (Figure 4A, B), as is evident from the presence of a bulge on the otherwise smooth descending half of the Y-arc. The presence of the $(C_3T_2)_{60}$ run in the lagging strand template does not result in a defined stall site, but rather leads to a widening of the Y-arc downstream from the repeat. Further experiments are needed to decipher the reasons for this arc widening (Figure 4A, B). Regardless, quantification of the stalling (Figure 4B, C, Supplementary Figure S7) reveals a significant increase in replication fork stalling with $(A_2G_3)_{60}$ in the lagging strand template over the flipped orientation (Unpaired t-test with Welch's correction, $P < 0.0212$) (Figure 4C). Similar orientation-dependence was previously observed for the Friedreich's ataxia $(GAA)_n$ repeat in many systems (42,55,56), which forms H-r triplexes *in vitro* (57–60) and in cells (28), suggesting the pathogenic CANVAS repeats may also form an H-r triplex in yeast cells.

In attempt to distinguish between H-DNA triplex- and G-quadruplex-mediated replication stalling, we knocked out the *PIF1* gene, which encodes a DNA helicase primarily implicated in the unwinding of G-quadruplexes. The absence of Pif1 had no significant effect on the strength of the stall, supporting a triplex-mediated stall (Supplementary Figure S8). Importantly, the replication fork does not stall when the nonpathogenic $(A_4G)_{60}$ repeat is in the lagging strand template, in line with our *in vitro* data showing minimal DNA polymerization stalling at the $(A_4G)_{10}$ repeat (Figure 1) and the lack of triplex formation by this repeat (Figure 2). Altogether, these data mirror our *in vitro* results in their orientation-dependence and pathogenic repeat-dependence.

Additional notable structures can be seen in the two-dimensional gels of $(A_2G_3)_{60}$ and $(A_4G)_{60}$ replication intermediates. One spot, observable directly above the 1.5n spot in the $(A_4G)_{60}$ sample is also seen with the no repeat control plas-

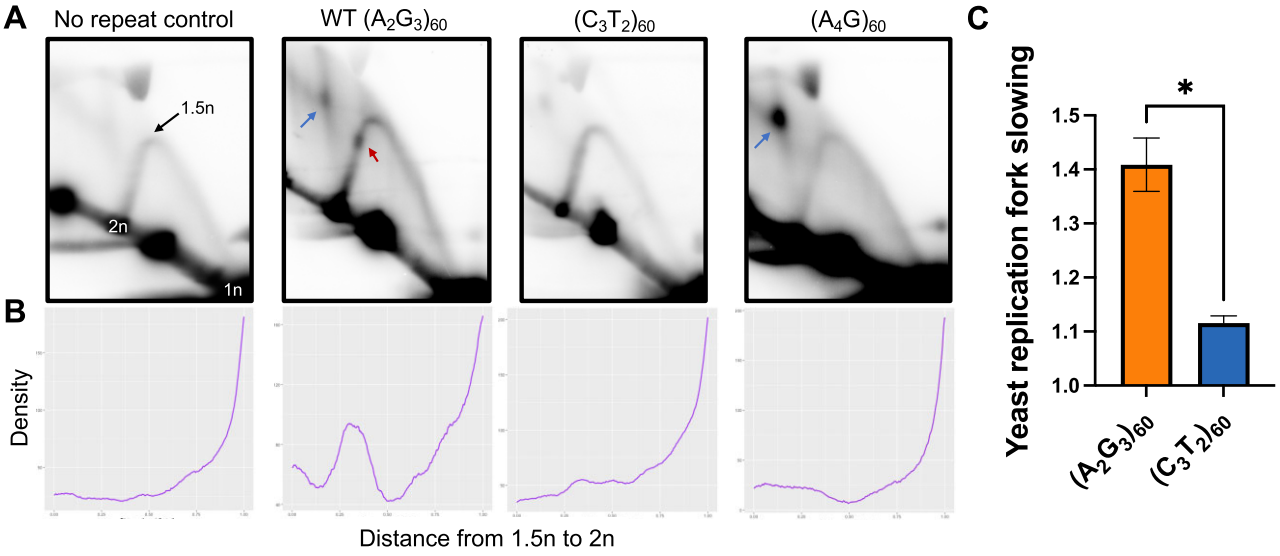


Figure 4. Analysis of yeast replication intermediates using two-dimensional gel electrophoresis demonstrates orientation-dependent stalling at $(A_2G_3)_{60}$ repeats. Details of replication intermediate collection and two-dimensional gel electrophoresis are described in Materials and Methods. Repeats were placed on the descending arm of the Y-arc. (**A**) Representative gels of the no repeat control, $(A_2G_3)_{60}$, $(C_3T_2)_{60}$ and $(A_4G)_{60}$ in the lagging strand template of replication in yeast from the yeast 2μ origin of replication. Three replicates were performed per repeat or orientation. The red arrow indicates replication fork stalling where the repeats are predicted to fall on the Y-arc. The blue arrow indicates the X-spot along the X-line. 1n, 1.5n and 2n mark replication intermediates at the beginning (1n), middle (1.5n) and end (2n) of replication. (**B**) Densitometry profiles along the arc starting from the 1.5n spot to the 2n spot. A custom script was used to plot these profiles (24). The density of the stall spot compared with the density along the Y-arc as well as the density of the background above and below the arc were graphed and used for quantification as described in Materials and Methods and shown in Supplementary Figure S7. (**C**) Quantification of replication fork slowing via area analysis. Error bars represent standard error of the mean. The signal of the stall over the arc intensity was compared using unpaired *t*-test with Welch's correction, $P < 0.0212$. Graph and statistical analysis performed using prism.

mid and can also be seen in other replicates of both $(A_4G)_{60}$ and the no repeat control. Since this spot is observed in the no repeat control samples, we conclude that it is not related to repeat-mediated replication stalling. In contrast, an X-line and a spot along this line are visible in $(A_2G_3)_{60}$, very prominent in $(A_4G)_{60}$, weakly present in the $(C_3T_2)_{60}$, and not present in the no repeat control intermediates (Figure 4A). The molecules on the X-line may be recombination intermediates or hemicatenanes, as was previously discussed for $(CTG)_n$ repeats (61). The peculiar spot roughly in the middle of the X-line may be indicative of a structure unable to branch migrate, as was previously observed for crosslinked recombination intermediates (62). We believe that this spot, as well as the X-line, are unrelated to the replication fork stalling at the repeat, given their prominence in the nonpathogenic repeat intermediates that do not display any fork stalling. Most likely, they reflect recombination intermediates, implying that a nonpathogenic, DUE-like $(A_4G)_{60}$ repeat is particularly prone for recombination. Further studies are needed to determine the nature of these structures.

Pathogenic $(A_2G_3)_n$ repeats stall replication in human cells in an orientation-dependent manner

Does the replication fork stalling by $(A_2G_3)_n$ repeats in yeast hold true in human cells? To answer this question, we utilized an episome replicating in human HEK293T cells described by us earlier (23). Briefly, this episome contains both the SV40 origin of replication and the T-antigen driving its replication initiation and elongation. Thus, the more it replicates, the more T-antigen is produced, driving subsequent rounds of replication. As a result, this system generates a high amount of replication intermediates, making their electrophoretic analysis feasible and relatively easy. The $(A_2G_3)_{60}$ repeat was cloned into this plasmid in two orientations relative to the SV40 origin.

Replication fork stalling patterns caused by the pathogenic $(A_2G_3)_n$ repeats in human cells appear to be fundamentally similar to that in yeast (Figure 5). When in the lagging strand template, the $(A_2G_3)_{60}$ run causes a very prominent replication stall on the ascending half of the Y-arc, while no distinct stall is detected in the no repeat control (Figure 5A, B). When the orientation of the repeats is flipped, there is significantly less fork stalling (Welch's ANOVA, $P < 0.0005$) (Figure 5A–C). Although there is increased density along the Y-arc when $(A_4G)_{60}$ serves as the template, this slight stalling is significantly less than that of $(A_2G_3)_{60}$ (Welch's ANOVA, $P < 0.0016$) (Figure 5A–C).

Given the intense stalling caused by the $(A_2G_3)_{60}$ repeats, we were interested in how the replication fork responds to try to overcome this obstacle. By placing the repeats on the descending arm of the Y arc, we can visualize replication fork intermediates that are otherwise unobservable when the repeats are on the ascending arm (23). This placement reveals additional intermediates that appear as a bulge above the stall site, as well as an X-spike that ultimately coalesces into a cone (Figure 6). Evident by their slower migration pattern in the second dimension, these intermediates are likely branched joint molecules (61) that arise in response to the prominent fork stalling. Whereas the cone represents terminal replication intermediates like fork convergence, the proximity of the bulge to the stall site indicates branched intermediates, likely resulting from fork reversal. While more experiments are required to determine the nature of these molecules, we hypothesize

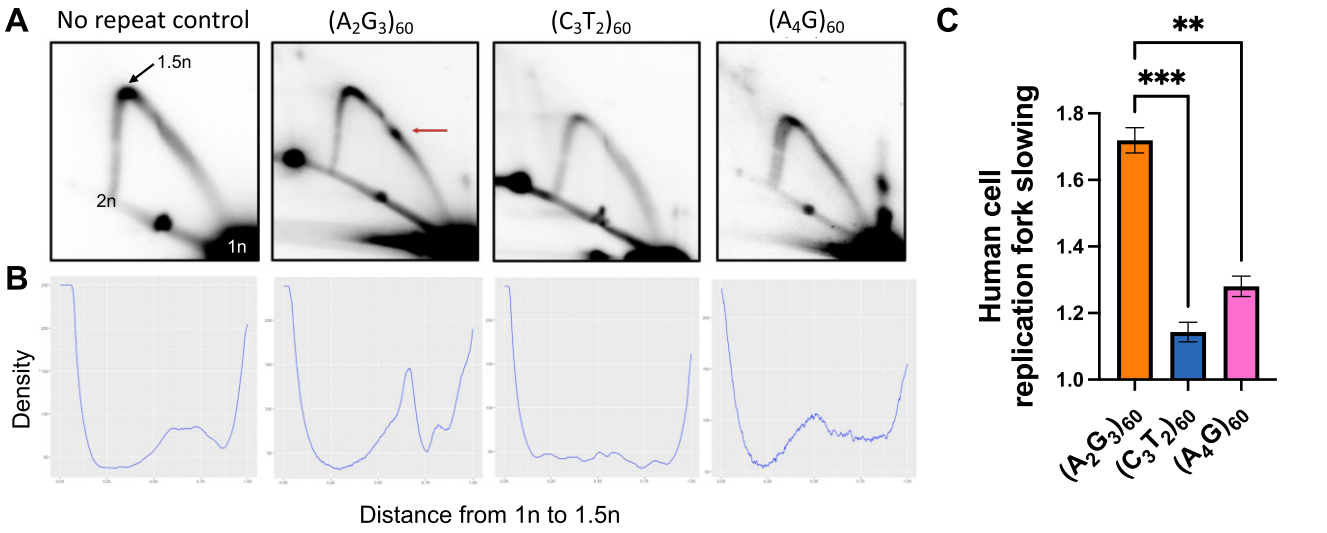


Figure 5. Analysis of human cell replication intermediates using two-dimensional gel electrophoresis demonstrates orientation-dependent stalling at $(A_2G_3)_{60}$ repeats. Details of replication intermediate collection and two-dimensional gel electrophoresis are described in Materials and methods. Repeats are placed on the ascending arm of the Y-arc. (**A**) Representative gels of the no repeat control, $(A_2G_3)_{60}$, $(C_3T_2)_{60}$ and $(A_4G)_{60}$ in the lagging strand template of replication from the SV40 origin of replication. Three replicates were performed per repeat orientation. The red arrow indicates replication fork stalling where the repeats are predicted to fall on the Y-arc. 1n, 1.5n and 2n mark replication intermediates at the beginning (1n), middle (1.5n) and end (2n) of replication. (**B**) Densitometry profiles along the arc starting at the 1n spot to the 1.5n spot. A custom script was used to plot these profiles (24). The density of the stall spot compared with the density along the Y-arc as well as the density of the background above and below the arc were graphed and used for quantification as described in Materials and Methods and shown in Supplementary Figure S7. (**C**) Quantification of replication fork slowing via area analysis. Error bars represent standard error of the mean. The signal of the stall over the arc intensity was compared using Welch's ANOVA test with Dunnett's multiple comparisons, $(A_2G_3)_{60}$ versus $(C_3T_2)_{60}$ $P < 0.0005$, $(A_2G_3)_{60}$ versus $(A_4G)_{60}$ $P < 0.0016$. Graph and statistical analysis performed using prism.

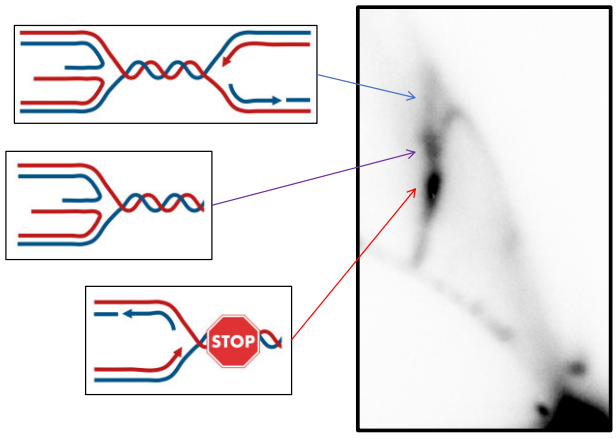


Figure 6. Analysis of human cell replication intermediates using two-dimensional gel electrophoresis demonstrates fork reversal at $(A_2G_3)_{60}$ repeats. Details of replication intermediate collection and two-dimensional gel electrophoresis are described in Materials and Methods. Repeats are placed on the descending arm of the Y-arc. Representative gel of $(A_2G_3)_{60}$ in the lagging strand template of replication from the SV40 origin of replication. Red arrow = replication fork stalling replication intermediates at the repeats on the Y-arc. Purple arrow = replication fork reversal intermediates. Blue arrow = cone structure with converging replication fork intermediates.

that fork reversal is utilized to overcome the $(A_2G_3)_{60}$-induced stall.

Discussion

CANVAS is a recently discovered neurodegenerative RED characterized by a spectrum of clinical manifestations, including but not limited to cerebellar ataxia, neuropathy, and vestibular areflexia (1,2). Though it is estimated to be the most common cause of hereditary late-onset ataxia (1,4), very little is known about its genetics and pathogenesis. An unusual feature of this RED is that a change in the sequence from the nonpathogenic $(A_4G)_{11–100}$ to the pathogenic $(A_2G_3)_{250–2000}$ causes CANVAS, rather than only the expansion of a repeat (1). No model systems have so far been developed to study CANVAS's genetics or pathogenesis. Clues to RED pathogenesis often lie within the repeats themselves, the non-B structures they form, and their interactions with cellular machinery. Using an arsenal of tools developed by our lab and others, we examined the CANVAS-causing $(A_2G_3)_n$ repeats and established key DNA-centric characteristics that may help unravel how these repeats expand and cause disease.

Importantly, the $(A_2G_3)_n$ repeat has the propensity to form an H-DNA triplex and a G-quadruplex. Based on our sequencing and primer extension experiments with various polymerases *in vitro*, the $(A_2G_3)_n$ repeat can stall polymerases by forming either of these structures, depending on ambient conditions. Meanwhile, $(A_4G)_{10}$ is not an obstacle to DNA polymerization (Figures 1, S1, S2).

Using chemical probing, we found that $(A_2G_3)_{10}$, but not $(A_4G)_{10}$, preferably forms an H-r DNA triplex in supercoiled DNA (Figure 2A). The S1-END-seq results generally support this conclusion, showing a much weaker propensity of $(A_4G)_n$ repeats to form triplexes genome-wide as compared to $(A_2G_3)_n$ repeats (Figure 3A). Importantly, more S1-END-seq peaks were detected at $(A_2G_3)_n$ repeats in replicating cells compared to G1-arrested cells (Figure 3B). The preferable formation of triplexes in cycling cells is in line with the strong, orientation-dependent replication fork stalling observed at the pathogenic repeat in yeast and human cells (Figures 4–6).

We observed replication fork stalling in yeast and human cells when the $(A_2G_3)_n$ repeat was in the lagging strand template of replication. H-motifs have been shown to stall replication when the homopurine strand is in the lagging strand template in bacteria (63), yeast (42,55,56), human cells, and human cell extracts (41). This replication fork stalling pattern is in stark contrast to the orientation-independent stalling induced by hairpin-forming sequences (64,65), while the pattern of replication fork stalling for G-quadruplexes is more complicated with evidence of G-quadruplexes impeding the replication fork when in the leading (66–68) or lagging (69) strand template. The Pif1 DNA helicase is known to unravel G4-mediated replication stalls in yeast (66,69). Furthermore, knocking out Pif1-helicase had no significant effect on replication stalling at $(A_2G_3)_n$ repeats in yeast (Supplementary Figure S8). Altogether, the CANVAS repeat's replication stalling is unlikely to be caused by G4-DNA. It is more consistent with a triplex formed by a single-stranded portion of the lagging strand template carrying the $(A_2G_3)_n$ run and the double-stranded repeat in front of the fork, similar to a model in Figure 3C. We hypothesize that triplex formation by the pathogenic repeat is induced by replication, simultaneously becoming an impediment for its progression through the repeats. This model is supported by the strong stall during both *in vitro* replication and replication in yeast and human cells and S1-END-seq mapping showing decreased triplex formation in non-replicating cells.

In the SV40-based plasmid system, we revealed additional replication intermediates indicative of fork reversal at the pathogenic $(A_2G_3)_n$ repeat. While the T-antigen plasmid replication system yields a large quantity of replication intermediates for analysis, there are major differences between this system and chromosomal replication machinery with regards to the protein composition. Specifically, in the SV40 replication fork, the large T-antigen replaces the CMG helicase and both leading and lagging DNA strands are synthesized by Pol δ DNA polymerase (70). Altogether, this makes the SV40 fork more prone for stalling and reversal, which can account for branched replication intermediates at the repeat-mediated stall observed in Figure 6. Future studies using the EBV (Epstein Barr virus)-EBNA1 replication system (71), which involves intact human replication fork machinery, would clarify if fork reversal at the $(A_2G_3)_n$ repeat is universal or specific for the SV40 replication fork.

Overall, we have found that the pathogenic $(A_2G_3)_n$ repeat forms a DNA triplex *in vitro* and stalls replication in an orientation-dependent manner both *in vitro* and in cells. In each experiment, we have juxtaposed the pathogenic $(A_2G_3)_n$ repeat with the nonpathogenic $(A_4G)_n$ repeat, finding the nonpathogenic repeat largely behaves similarly to a nonrepetitive sequence and thus does not impede replication to the extent that $(A_2G_3)_n$ does. Therefore, we believe we have uncovered features of the pathogenic allele that could be integral to its instability and pathogenicity. This illuminates an important next step: studying the genetics of CANVAS and how the repeats expand and cause disease. To this point, work to establish model systems to study the repeats' instability (contraction and expansion) is currently underway and is a crucial start to understanding this recently characterized disease. Similarly, conducting these structure-functional analysis experiments with additional pathogenic and nonpathogenic alleles may illuminate a pattern suggestive of one structure over another, though it is possible different structures are formed by different repeats, leading to the same disease.

Data availability

The data underlying this article are available in the article and in its online Supplementary material.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank Catherine Freudenreich, Mitch McVey, Claire Moore, Ralph Scully, Liangzi Li, members of the Freudenreich and McVey laboratories, and other members of the Mirkin laboratory for their integral input to this project.

Funding

The work in the Mirkin laboratory is supported by the National Institute of General Medical Sciences [R35GM130322]; National Science Foundation [2153071]; Ryan McGinty's work is supported by the National Institute of General Medical Sciences [R35GM127131]; Nussenzweig laboratory is supported by the Intramural Research Program of the NIH funded in part with federal funds from the NCI [HHSN2612015000031]; Ellison Medical Foundation Senior Scholar in Aging Award [AG-SS-2633-11]; Department of Defense Awards [W81XWH-16-1-599 and W81XWH-19-1-0652]; Alex's Lemonade Stand Foundation Award, an NIH Intramural FLEX Award; Friedreich's Ataxia Research Alliance. Funding for open access charge: NIGMS.

Conflict of interest statement

None declared.

References

1. Cortese,A., Simone,R., Sullivan,R., Vandrovcova,J., Tariq,H., Yau,W.Y., Humphrey,J., Jaunmuktane,Z., Sivakumar,P., Polke,J., *et al.* (2019) Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nat. Genet.*, **51**, 649–658.
2. Rafehi,H., Szmulewicz,D.J., Bennett,M.F., Sobreira,N.L.M., Pope,K., Smith,K.R., Gillies,G., Diakumis,P., Dolzhenko,E., Eberle,M.A., *et al.* (2019) Bioinformatics-based identification of expanded repeats: a non-reference intronic pentamer expansion in RFC1 causes CANVAS. *Am. J. Hum. Genet.*, **105**, 151–165.
3. Arteche-López,A., Avila-Fernandez,A., Damian,A., Soengas-Gonda,E., de la Fuente,R.P., Gómez,P.R., Merlo,J.G., Burgos,L.H., Fernández,C.C., Rosales,J.M.L., *et al.* (2023) New Cerebellar Ataxia, Neuropathy, Vestibular Areflexia Syndrome cases are caused by the presence of a nonsense variant in compound heterozygosity with the pathogenic repeat expansion in the RFC1 gene. *Clin. Genet.*, **103**, 236–241.
4. Cortese,A., Curro',R., Vegezzi,E., Yau,W.Y., Houlden,H. and Reilly,M.M. (2022) Cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS): genetic and clinical aspects. *Pract. Neurol.*, **22**, 14–18.
5. Cortese,A., Tozza,S., Yau,W.Y., Rossi,S., Beecroft,S.J., Jaunmuktane,Z., Dyer,Z., Ravenscroft,G., Lamont,P.J., Mossman,S., *et al.* (2020) Cerebellar ataxia, neuropathy, vestibular

areflexia syndrome due to RFC1 repeat expansion. *Brain*, **143**, 480–490.
6. Ogi,T., Limsirichaikul,S., Overmeer,R.M., Volker,M., Takenaka,K., Cloney,R., Nakazawa,Y., Niimi,A., Miki,Y., Jaspers,N.G., *et al.* (2010) Three DNA polymerases, recruited by different mechanisms, carry out NER repair synthesis in human cells. *Mol. Cell*, **37**, 714–727.
7. Iyama,T. and Wilson,D.M. (2013) DNA repair mechanisms in dividing and non-dividing cells. *DNA Repair (Amst.)*, **12**, 620–636.
8. Benkirane,M., Da Cunha,D., Marelli,C., Larrieu,L., Renaud,M., Varilh,J., Pointaux,M., Baux,D., Ardouin,O., Vangoethem,C., *et al.* (2022) RFC1 nonsense and frameshift variants cause CANVAS: clues for an unsolved pathophysiology. *Brain*, **145**, 3770–3775.
9. Ronco,R., Perini,C., Currò,R., Dominik,N., Facchini,S., Gennari,A., Simone,R., Stuart,S., Nagy,S., Vegezzi,E., *et al.* (2023) Truncating variants in RFC1 in cerebellar ataxia, neuropathy, and vestibular areflexia syndrome. *Neurology*, **100**, e543–e554.
10. King,K.A., Wegner,D.J., Bucelli,R.C., Shapiro,J., Paul,A.J., Dickson,P.I., Wambach,J.A. and and Undiagnosed Disease Network (UDN)and Undiagnosed Disease Network (UDN) (2022) Whole-genome and long-read sequencing identify a novel mechanism in RFC1 resulting in CANVAS syndrome. *Neurol. Genet*, **8**, e200036.
11. Weber,S., Coarelli,G., Heinzmann,A., Monin,M.-L., Richard,N., Gerard,M., Durr,A. and Huin,V. (2022) Two RFC1 splicing variants in CANVAS. *Brain*, **146**, e14–e16.
12. Khristich,A.N. and Mirkin,S.M. (2020) On the wrong DNA track: molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.*, **295**, 4134–4170.
13. Dominik,N., Magri,S., Currò,R., Abati,E., Facchini,S., Corbetta,M., MacPherson,H., Di Bella,D., Sarto,E., Stevanovski,I., *et al.* (2023) Normal and pathogenic variation of RFC1 repeat expansions: implications for clinical diagnosis. *Brain*, **146**, 5060–5069.
14. Barghigiani,M., De Michele,G., Tessa,A., Fico,T., Natale,G., Saccà,F., Pane,C., Cuomo,N., De Rosa,A., Pappatà,S., *et al.* (2022) Screening for RFC-1 pathological expansion in late-onset ataxias: a contribution to the differential diagnosis. *J. Neurol.*, **269**, 5431–5435.
15. Akçimen,F., Ross,J.P., Bourassa,C.V., Liao,C., Rochefort,D., Gama,M.T.D., Dicaire,M.-J., Barsottini,O.G., Brais,B., Pedroso,J.L., *et al.* (2019) Investigation of the RFC1 repeat expansion in a Canadian and a Brazilian Ataxia cohort: identification of novel conformations. *Front. Genet.*, **10**, 1219.
16. Abramzon,Y., Dewan,R., Cortese,A., Resnick,S., Ferrucci,L., Houlden,H. and Traynor,B.J. (2021) Investigating RFC1 expansions in sporadic amyotrophic lateral sclerosis. *J. Neurol. Sci.*, **430**, 118061.
17. Scriba,C.K., Beecroft,S.J., Clayton,J.S., Cortese,A., Sullivan,R., Yau,W.Y., Dominik,N., Rodrigues,M., Walker,E., Dyer,Z., *et al.* (2020) A novel RFC1 repeat motif (ACAGG) in two Asia-Pacific CANVAS families. *Brain*, **143**, 2904–2910.
18. Erdmann,H., Schöberl,F., Giurgiu,M., Leal Silva,R.M., Scholz,V., Scharf,F., Wendlandt,M., Kleinle,S., Deschauer,M., Nübling,G., *et al.* (2022) Parallel in-depth analysis of repeat expansions in ataxia patients by long-read sequencing. *Brain*, **146**, 1831–1843.
19. Mirkin,S.M. and Frank-Kamenetskii,M.D. (1994) H-DNA and related structures. *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 541–576.
20. Ding,Y., Fleming,A.M. and Burrows,C.J. (2018) Case studies on potential G-quadruplex-forming sequences from the bacterial orders Deinococcales and Thermales derived from a survey of published genomes. *Sci. Rep.*, **8**, 15679.
21. Shah,K.A., Shishkin,A.A., Voineagu,I., Pavlov,Y.I., Shcherbakova,P.V. and Mirkin,S.M. (2012) Role of DNA polymerases in repeat-mediated genome instability. *Cell Rep.*, **2**, 1088–1095.
22. Neil,A.J., Hisey,J.A., Quasem,I., McGinty,R.J., Hitczenko,M., Khristich,A.N. and Mirkin,S.M. (2021) Replication-independent instability of Friedreich's ataxia GAA repeats during chronological aging. *Proc. Natl Acad. Sci. U.S.A.*, **118**, e2013080118.
23. Rastokina,A., Cebrián,J., Mozafari,N., Mandel,N.H., Smith,C.I.E., Lopes,M., Zain,R. and Mirkin,S.M. (2023) Large-scale expansions of Friedreich's ataxia GAA•TTC repeats in an experimental human system: role of DNA replication and prevention by LNA-DNA oligonucleotides and PNA oligomers. *Nucleic Acids Res.*, **51**, 8532–8549.
24. Radchenko,E.A., Aksenova,A.Y., Volkov,K.V., Shishkin,A.A., Pavlov,Y.I. and Mirkin,S.M. (2022) Partners in crime: tbf1 and Vid22 promote expansions of long human telomeric repeats at an interstitial chromosome position in yeast. *PNAS Nexus*, **1**, pgac080.
25. Krasilnikov,A.S., Panyutin,I.G., Samadashwily,G.M., Cox,R., Lazurkin,Y.S. and Mirkin,S.M. (1997) Mechanisms of triplex-caused polymerization arrest. *Nucleic Acids Res.*, **25**, 1339–1346.
26. Gardner,A.F. and Jack,W.E. (1999) Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. *Nucleic Acids Res.*, **27**, 2545–2553.
27. Hernandez,A.J., Lee,S.-J., Chang,S., Lee,J.A., Loparo,J.J. and Richardson,C.C. (2020) Catalytically inactive T7 DNA polymerase imposes a lethal replication roadblock. *J. Biol. Chem.*, **295**, 9542–9550.
28. Matos-Rodrigues,G., van Wietmarschen,N., Wu,W., Tripathi,V., Koussa,N.C., Pavani,R., Nathan,W.J., Callen,E., Belinky,F., Mohammed,A., *et al.* (2022) S1-END-seq reveals DNA secondary structures in human cells. *Mol. Cell*, **82**, 3538–3552.
29. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
30. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and and 1000 Genome Project Data Processing Subgroupand 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
31. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
32. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
33. McGinty,R.J. and Sunyaev,S.R. (2023) Revisiting mutagenesis at non-B DNA motifs in the human genome. *Nat. Struct. Mol. Biol.*, **30**, 417–424.
34. Hänsel-Hertsch,R., Beraldi,D., Lensing,S.V., Marsico,G., Zyner,K., Parry,A., Di Antonio,M., Pike,J., Kimura,H., Narita,M., *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267–1272.
35. Krasilnikova,M.M. and Mirkin,S.M. (2004) Analysis of triplet repeat replication by two-dimensional gel electrophoresis. In: Kohwi,Y. (ed.) *Trinucleotide Repeat Protocols, Methods in Molecular Biology™*. Humana Press, Totowa, NJ, pp. 19–28.
36. Voineagu,I., Freudenreich,C.H. and Mirkin,S.M. (2009) Checkpoint responses to unusual structures formed by DNA repeats. *Mol. Carcinog.*, **48**, 309–318.
37. Samadashwily,G.M., Raca,G. and Mirkin,S.M. (1997) Trinucleotide repeats affect DNA replication in vivo. *Nat. Genet.*, **17**, 298–304.
38. Hile,S.E. and Eckert,K.A. (2004) Positive correlation between DNA polymerase alpha-primase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences. *J. Mol. Biol.*, **335**, 745–759.
39. Anand,R.P., Shah,K.A., Niu,H., Sung,P., Mirkin,S.M. and Freudenreich,C.H. (2012) Overcoming natural replication barriers: differential helicase requirements. *Nucleic Acids Res.*, **40**, 1091–1105.

40. Patel,H.P., Lu,L., Blaszak,R.T. and Bissler,J.J. (2004) PKD1 intron 21: triplex DNA formation and effect on replication. *Nucleic Acids Res.*, **32**, 1460–1468.
41. Liu,G., Myers,S., Chen,X., Bissler,J.J., Sinden,R.R. and Leffak,M. (2012) Replication fork stalling and checkpoint activation by a PKD1 locus mirror repeat polypurine-polypyrimidine (Pu-Py) tract. *J. Biol. Chem.*, **287**, 33412–33423.
42. Krasilnikova,M.M. and Mirkin,S.M. (2004) Replication stalling at Friedreich's ataxia (GAA)n repeats in vivo. *Mol. Cell. Biol.*, **24**, 2286–2295.
43. Masnovo,C., Lobo,A.F. and Mirkin,S.M. (2022) Replication dependent and independent mechanisms of GAA repeat instability. *DNA Repair (Amst.)*, **118**, 103385.
44. Wang,G. and Vasquez,K.M. (2022) Dynamic alternative DNA structures in biology and disease. *Nat. Rev. Genet.*, **24**, 211–234.
45. Vander Horn,P.B., Davis,M.C., Cunniff,J.J., Ruan,C., McArdle,B.F., Samols,S.B., Szasz,J., Hu,G., Hujer,K.M., Domke,S.T., *et al.* (1997) Thermo Sequenase DNA polymerase and T. acidophilum pyrophosphatase: new thermostable enzymes for DNA sequencing. *BioTechniques*, **22**, 758–762.
46. Bhattacharyya,D., Mirihana Arachchilage,G. and Basu,S. (2016) Metal cations in G-quadruplex folding and stability. *Front Chem*, **4**, 38.
47. Chashchina,G.V., Beniaminov,A.D. and Kaluzhny,D.N. (2019) Stable G-quadruplex structures of oncogene promoters induce potassium-dependent stops of thermostable DNA polymerase. *Biochemistry*, **84**, 562–569.
48. Castillo Bosch,P., Segura-Bayona,S., Koole,W., van Heteren,J.T., Dewar,J.M., Tijsterman,M. and Knipscheer,P. (2014) FANCJ promotes DNA synthesis through G-quadruplex structures. *EMBO J.*, **33**, 2521–2533.
49. Malkov,V.A., Voloshin,O.N., Soyfer,V.N. and Frank-Kamenetskii,M.D. (1993) Cation and sequence effects on stability of intermolecular pyrimidine-purine-purine triplex. *Nucleic Acids Res.*, **21**, 585–591.
50. Lee,S.-J. and Richardson,C.C. (2011) Choreography of bacteriophage T7 DNA replication. *Curr. Opin. Chem. Biol.*, **15**, 580–586.
51. Frank-Kamenetskii,M.D. and Mirkin,S.M. (1995) Triplex dna structures. *Annu. Rev. Biochem.*, **64**, 65–95.
52. Aymami,J., Coll,M., Frederick,C.A., Wang,A.H. and Rich,A. (1989) The propeller DNA conformation of poly(dA).Poly(dT). *Nucleic Acids Res.*, **17**, 3229–3245.
53. Kowalski,D. and Eddy,M.J. (1989) The DNA unwinding element: a novel, cis-acting component that facilitates opening of the Escherichia coli replication origin. *EMBO J.*, **8**, 4335–4344.
54. Rubin,C.M. and Schmid,C.W. (1980) Pyrimidine-specific chemical reactions useful for DNA sequencing. *Nucleic Acids Res.*, **8**, 4613–4619.
55. Kim,H.-M., Narayanan,V., Mieczkowski,P.A., Petes,T.D., Krasilnikova,M.M., Mirkin,S.M. and Lobachev,K.S. (2008) Chromosome fragility at GAA tracts in yeast depends on repeat orientation and requires mismatch repair. *EMBO J.*, **27**, 2896–2906.
56. Shishkin,A.A., Voineagu,I., Matera,R., Cherng,N., Chernet,B.T., Krasilnikova,M.M., Narayanan,V., Lobachev,K.S. and Mirkin,S.M. (2009) Large-scale expansions of Friedreich's Ataxia GAA repeats in yeast. *Mol. Cell*, **35**, 82–92.
57. Gacy,A.M., Goellner,G.M., Spiro,C., Chen,X., Gupta,G., Bradbury,E.M., Dyer,R.B., Mikesell,M.J., Yao,J.Z., Johnson,A.J., *et al.* (1998) GAA instability in Friedreich's Ataxia shares a common, DNA-directed and intraallelic mechanism with other trinucleotide diseases. *Mol. Cell*, **1**, 583–593.
58. Mariappan,S.V., Catasti,P., Silks,L.A., Bradbury,E.M. and Gupta,G. (1999) The high-resolution structure of the triplex formed by the GAA/TTC triplet repeat associated with Friedreich's ataxia. *J. Mol. Biol.*, **285**, 2035–2052.
59. Sakamoto,N., Chastain,P.D., Parniewski,P., Ohshima,K., Pandolfo,M., Griffith,J.D. and Wells,R.D. (1999) Sticky DNA: self-association properties of long GAA·TTC repeats in R·R·Y triplex structures from Friedreich's ataxia. *Mol. Cell*, **3**, 465–475.
60. Vetcher,A.A., Napierala,M., Iyer,R.R., Chastain,P.D., Griffith,J.D. and Wells,R.D. (2002) Sticky DNA, a long GAA.GAA.TTC triplex that is formed intramolecularly, in the sequence of intron 1 of the frataxin gene. *J. Biol. Chem.*, **277**, 39217–39227.
61. Nguyen,J.H.G., Viterbo,D., Anand,R.P., Verra,L., Sloan,L., Richard,G.-F. and Freudenreich,C.H. (2017) Differential requirement of Srs2 helicase and Rad51 displacement activities in replication of hairpin-forming CAG/CTG repeats. *Nucleic Acids Res.*, **45**, 4519–4531.
62. Giannattasio,M., Zwicky,K., Follonier,C., Foiani,M., Lopes,M. and Branzei,D. (2014) Visualization of recombination-mediated damage bypass by template switching. *Nat. Struct. Mol. Biol.*, **21**, 884–892.
63. Pollard,L.M., Sharma,R., Gómez,M., Shah,S., Delatycki,M.B., Pianese,L., Monticelli,A., Keats,B.J.B. and Bidichandani,S.I. (2004) Replication-mediated instability of the GAA triplet repeat mutation in Friedreich ataxia. *Nucleic Acids Res.*, **32**, 5962–5971.
64. Pelletier,R., Krasilnikova,M.M., Samadashwily,G.M., Lahue,R. and Mirkin,S.M. (2003) Replication and expansion of trinucleotide repeats in yeast. *Mol. Cell. Biol.*, **23**, 1349–1357.
65. Liu,G., Chen,X., Gao,Y., Lewis,T., Barthelemy,J. and Leffak,M. (2012) Altered replication in human cells promotes DMPK (CTG)(n) · (CAG)(n) repeat instability. *Mol. Cell. Biol.*, **32**, 1618–1632.
66. Lopes,J., Piazza,A., Bermejo,R., Kriegsman,B., Colosio,A., Teulade-Fichou,M.-P., Foiani,M. and Nicolas,A. (2011) G-quadruplex-induced instability during leading-strand replication. *EMBO J.*, **30**, 4033–4046.
67. Guilbaud,G., Murat,P., Recolin,B., Campbell,B.C., Maiter,A., Sale,J.E. and Balasubramanian,S. (2017) Local epigenetic reprogramming induced by G-quadruplex ligands. *Nat. Chem.*, **9**, 1110–1117.
68. Sarkies,P., Murat,P., Phillips,L.G., Patel,K.J., Balasubramanian,S. and Sale,J.E. (2012) FANCJ coordinates two pathways that maintain epigenetic stability at G-quadruplex DNA. *Nucleic Acids Res.*, **40**, 1485–1498.
69. Dahan,D., Tsirkas,I., Dovrat,D., Sparks,M.A., Singh,S.P., Galletto,R. and Aharoni,A. (2018) Pif1 is essential for efficient replisome progression through lagging strand G-quadruplex DNA secondary structures. *Nucleic Acids Res.*, **46**, 11847–11857.
70. Sowd,G.A. and Fanning,E. (2012) A wolf in sheep's clothing: SV40 Co-opts host genome maintenance proteins to replicate viral DNA. *PLoS Pathog.*, **8**, e1002994.
71. Willis,N.A., Chandramouly,G., Huang,B., Kwok,A., Follonier,C., Deng,C. and Scully,R. (2014) BRCA1 controls homologous recombination at Tus/Ter-stalled mammalian replication forks. *Nature*, **510**, 556–559.

Received: July 25, 2023. **Revised:** February 6, 2024. **Editorial Decision:** February 6, 2024. **Accepted:** February 8, 2024