A study of the mechanisms of GAA repeat expansion in *S. cerevisiae*

By Miranda U. Liang

A Senior Honors Thesis for the Department of Biology
Tufts University, 2018

Committee: Dr. Sergei Mirkin, Dr. Catherine Freudenreich, Dr. Alex Neil

Acknowledgments

I would like to thank the famous Dr. Sergei Mirkin, not only for giving me this amazing opportunity to work in his lab, but also for his wonderful guidance and mentorship. I have found his wisdom about science and also life in general to be incredibly valuable. He has been such an approachable and patient teacher who is always happy to explain something to me. I will never forget a single one of his many entertaining stories.

I would like to thank Dr. Alex Neil for being the best mentor ever. He has taught me how to be a truly independent researcher. Every step of the way, he has encouraged me but also challenged me to think for myself and draw my own conclusions. I aspire to one day be as great a scientist as Alex. He has also been a great lab companion who gives good music recommendations and is happy to talk philosophy with me.

I would like to thank Dr. Catherine Freudenreich for serving on my committee and for her invaluable expertise and suggestions. I would also like to thank my biology advisor Dr. Juliet Fuhrman for being super encouraging and helpful. I am grateful to the National Science Foundation and the REU program for funding my research the summer of 2017.

I would like to thank all past and current members of the Mirkin Lab for making the lab such a nice place to work: Dr. Jane Kim, Dr. Elina Radchenko, Dr. Jorge Cebrian Castillo, Dr. Artem Kononenko, Katia Spivakovsky-González, Sasha Khristich, Teresa Dinter, Tri Hoang, Clara Williamson, Jessica Maya, Sarah Burton, Grace Materne, and Anna Kolchinski. Special shoutout to Dr. Ryan McGinty who read over several of my research papers even when I asked really late notice and to Julia Hisey, Rachel Rubinstein, and Tayaba Naz for making summers at the lab especially fun.

I would like to thank my parents for their love and support. I am also grateful for the home-cooked meals they would send me every week when school got too hard. I would also like to thank Hannah Harris, my lunch buddy, and my friends for their constant encouragement, including but not limited to, Abriana, Carolina, Caroline, Julie, Lydia, Maya, and Ming.

Thanks to all of you!

Table of Contents

Chapter 1 - Repeat orientation and alternate DNA structures

Abstract

Friedreich's ataxia (FRDA) is a genetic, neurodegenerative disease caused by the inheritance of abnormally expanded GAA repeats in the first intron of the frataxin gene (*FXN*). Disease severity increases with longer expansions, making the mechanism of GAA repeat expansion of great interest. Past research has implicated transcription-replication collisions and R-loop (RNA/DNA hybrid) formation as possible instigators of GAA repeat expansion. To study these phenomena, we cloned three cassettes that allow for the measurement of GAA repeat expansion rates in *S. cerevisiae*. These cassettes differed in three variables: whether the GAA repeats were located on the replicative leading or lagging strand template, whether the sense strand of transcription contained GAA or UUC repeats, and whether transcription across the repeats was oriented co-directionally or head-on with replication. We also knocked out the RNase H1 and RNase H201 enzymes that degrade R-loops in strains bearing our three cassettes and measured GAA expansion rate. Our results indicated that transcription-induced expansion and R-loop induced expansion are orientation-independent. To explain our results, we propose that transcribed GAA repeats have the propensity to form into an H-loop, a structure that combines a DNA triplex (H-DNA) with an R-loop, which blocks replication and ultimately results in repeat expansion.

Introduction

DNA microsatellites and human disease

DNA microsatellites are sequences composed of short tandem repeats (1-13 bp units) that make up approximately 3% of the human genome (Lander et al. 2001). They have been observed in all parts of the genome, including exons, introns, and intergenic regions (Subramanian et al. 2003).

Microsatellites are of particular interest because they are prone to length polymorphisms (Sutherland & Richards 1995). Depending on the length and type of microsatellite, mutation rate can be as high as 10^{-2} mutations per locus per generation, many orders of magnitude higher than the rate of single base mutations (Ellegren 2004). These length polymorphisms can be used for many practical purposes, including: cancer diagnoses, paternity/maternity testing, forensic fingerprinting, genetic linkage analysis, and population genetics analysis.

In some cases, these polymorphisms are deleterious. Almost 30 human genetic diseases are caused by the expansion of microsatellites, including Huntington's disease (CAG repeats), myotonic dystrophy type 2 (CCTG repeats), and Friedreich's ataxia (GAA repeats) (Zhao & Usdin 2015). In addition, more disease-causing microsatellites continue to be identified to this day. For example, expanded GGGGCC repeats in the c9orf72 gene were recently discovered to be a major cause of familial amyotrophic lateral sclerosis (ALS) and familial frontotemporal dementia (FTD) (DeJesus-Hernandez et al. 2011). All of these microsatellites are capable of forming secondary DNA structures, such as hairpins, H-DNA (triplex DNA), and G4 quadruplexes (Mirkin 2007, Zhao & Usdin 2015).

The expansion of microsatellites becomes more frequent the longer the repetitive sequence becomes (Mirkin 2007). Consequently, many of microsatellite-associated diseases exhibit genetic anticipation, a phenomenon where disease severity increases and age of onset decreases with every successive generation.

Friedreich's ataxia and GAA repeats

Friedreich's ataxia (FRDA) is an autosomal recessive, neurodegenerative disease caused by the abnormal expansion of GAA repeats in the first intron of the frataxin gene (*FXN*) (Campuzano et al. 1996). The most frequently inherited form of ataxia, the disease has a prevalence of 1:50,000 in European populations and causes symptoms that include ataxia (a lack of voluntary muscle coordination), muscle weakness, and hypertrophic cardiomyopathy (Campuzano et al. 1996). Frataxin is an enzyme that participates in iron-sulfur cluster biosynthesis in the mitochondria (Lupoli et al. 2017). When the number of GAA repeats at the *FXN* locus expand to disease-length, it results in a loss of function due to heterochromatin formation at the repeats causing transcriptional silencing of the *FXN* gene (Li et al. 2015).

Disease-causing alleles have been observed to contain between 70 to over 1000 GAA repeats (Pandolfo 2001). Alleles containing GAA repeats that are pre-mutation length (30-70 repeats) or disease-length can expand throughout a person's lifetime (somatically) and between generations (meiotically) (Cossée et al. 1997, Sharma et al. 2002, Filla et al. 1996). FRDA does not demonstrate genetic anticipation because it is an autosomal recessive disease that will skip generations in most cases. However, it is still true that increased repeat length has been correlated with disease severity and inversely correlated with age of disease onset (Montermini et al. 1997). Thus, the mechanism by which repeats expand is of great interest and may provide insight into the pathophysiology of the disease.

To study repeat expansion, we use *Saccharomyces cerevisiae* (baker's yeast) in our experiments. Yeast is an excellent model organism for genetics studies, as they replicate quickly, and their genomes are simple to manipulate, allowing the selection of genetic traits. Most importantly, the mechanisms of DNA replication and repair are well conserved from humans to yeast.

Orientation-dependence of GAA repeat expansion

The rate of microsatellite instability has been observed to be dependent on repeat orientation with respect to the direction of DNA replication. For example, a high rate of instability

is observed when the lagging strand template contains CTG repeats, and the complementary CAG repeats are on the leading strand template (Kovtun & McMurray 2008). However, a lower rate is observed in the opposite orientation, with CAG repeats on the lagging strand template and CTG repeats on the leading strand template (Kovtun & McMurray 2008).

Instability of GAA repeats also demonstrates orientation-dependence. In yeast, when GAA repeats were located on the lagging strand template rather than the leading strand template, there was more replication fork stalling and a higher rate of expansion (Shishkin et al. 2009). In addition, GAA repeats were shown to induce higher rates of gross chromosomal rearrangements and DNA breakage when located on the lagging strand template rather than the leading strand template (Kim et al. 2008). Lastly, expanded GAA repeats were shown to stall replication forks at the endogenous FXN locus in human stem cells and to have a preferred orientation of replication such that the GAA repeats are located on leading strand template (Gerhardt et al. 2016).

For our study, we utilized genetic constructs in yeast that contain GAA repeats in both orientations with respect to replication. We further varied the orientation of the repeats with respect to transcription. This allowed us to distinguish the contributions of replication and transcription orientations as they relate to two phenomena: transcription-replication collisions and R-loops.

Transcription-replication collisions

Genes can either be oriented towards or away from the incoming replication forks. Therefore, collisions between replication and transcription can either occur head-on or co-directionally. At the endogenous FXN locus, replication was observed to move in both directions across the gene (co-directionally and head-on with transcription) (Gerhardt et al. 2016).

We were interested in the possible effects that conflicts between replication and transcription machinery could have on GAA expansion rate. In general, head-on collisions between replication and transcription have been shown to be more mutagenic than co-

directional collisions, such that it has been theorized that bacteria have evolved to orient most of their genes co-directionally (Rocha & Danchin 2003). Furthermore, several studies have linked head-on collisions to homologous recombination (Takeuchi et al. 2003, Prado & Aguilera, 2005) and trinucleotide repeat instability (Rindler & Bidichandani, 2010). In contrast, another study in yeast observed a higher rate of replication fork stalling and a higher expansion rate in yeast when the transcription of GAA repeats were oriented co-directionally rather than head-on with replication (Shishkin et al. 2009)

We hypothesized that head-on collisions between replication and transcription could contribute to GAA repeat expansion. To investigate the relationship between transcription-replication collision orientation and repeat expansion, we compared GAA expansion rate between cassettes in which replication and transcription were moving in either a head-on or a co-directional direction.

R-loop formation across GAA repeats

An R-loop is a structure that can form during transcription when the nascent mRNA strand binds to the template strand of DNA and displaces the non-template strand. R-loops play roles in several biological processes, including transcription termination, mitochondrial DNA replication, and immunoglobulin class-switch recombination (Groh & Gromak 2014).

However, R-loops can also have deleterious effects. R-loop formation has been shown to stall replication fork progression (Gan et al., 2011), block transcription (Belotserkovskii et al. 2012), and contribute to genome instability by inducing DNA double strand breaks, transcription-associated recombination, and chromosome loss (Groh & Gromak 2014).

R-loops are prone to forming in G-rich, homopurine stretches of DNA (Belotserkovskii et al. 2012). *In vitro*, duplexes between purine-rich RNA and pyrimidine-rich DNA strands were shown to be stronger than those between pyrimidine-rich RNA and purine-rich DNA strands or between two DNA strands (Gyi et al. 1996). Thus, RNA strands containing GAA repeats would theoretically be prone to R-loop formation, and indeed, it has been shown that R-loops form

across expanded GAA repeats at the endogenous FXN locus and impede transcription of the gene (Groh et al. 2014).

There is evidence that R-loop formation contributes to GAA repeat expansion. Increased transcription of GAA repeats has been shown to increase the rate of expansion, possibly due to increased R-loop formation (Shah et al. 2014). In addition, R-loops have been shown to induce instability of CAG repeats in yeast (Su & Freudenreich 2017). Lastly, in strains lacking the RNase H enzymes that degrade R-loops (RNase H1 and RNase H201), GAA expansion rates are significantly elevated, but only when the repeats are transcribed (Neil et al. 2018).

To investigate the relationship between R-loop formation and repeat expansion, we compared cassettes in which either GAA or TTC repeats are transcribed into RNA. We hypothesized that an RNA strand containing the pyrimidine-rich TTC repeats would be less conducive for R-loop formation and therefore would result in a lower rate of R-loop-mediated repeat expansion.

Goals of this project

The goal of this project was to explore the orientation-dependence of GAA repeat expansion with respect to both transcription and replication. We accomplished this using several selectable gene cassettes containing 100 GAA repeats that detected expansion events. We began with the GAA;CD cassette, wherein transcription and replication are oriented co-directionally, and the GAA repeats are located on the replicative lagging strand template, as well as on the transcriptional sense strand.

To investigate the effect of transcription-replication orientation on expansion, we inverted the GAA;CD cassette with respect to the origin of replication, creating the GAA;HO cassette, and then compared the two. However, by inverting the GAA;CD cassette, we also modified the lagging strand template to contain TTC repeats instead of GAA repeats. Therefore, we also created a TTC;CD cassette, wherein only the GAA repeats are inverted with respect to the origin of replication, to act as a control.

Analogously, to investigate the effect of RNA composition on R-loop formation and repeat expansion, we compared expansion rates of the GAA;CD and TTC;CD cassettes. However, because inverting the GAA repeats into TTC repeats also modified the lagging strand template to contain TTC repeats instead of GAA repeats, the GAA;HO cassette acted as a control in this case.

Ultimately, we observed that a co-directional, not head-on, orientation between replication and transcription promoted more repeat expansions. We also observed that R-loop-mediated repeat expansion occurred independently of the identity of the repeats on the transcriptional sense strand. Based on these observations, we formulated an orientation-independent model for R-loop-mediated repeat expansion wherein the GAA repeats form an H-loop—a structure that combines triplex DNA with an R-loop.

Materials & Methods

Yeast strains

The *S. cerevisiae* strains used in these experiments were derived from the parent strain CH1585 (*MAT***a**, *leu2-Δ1, trp1-Δ63, ura3-52, his3-200*). Knockout strains were created with gene replacement and verified by PCR with external and internal primers to the deleted gene.

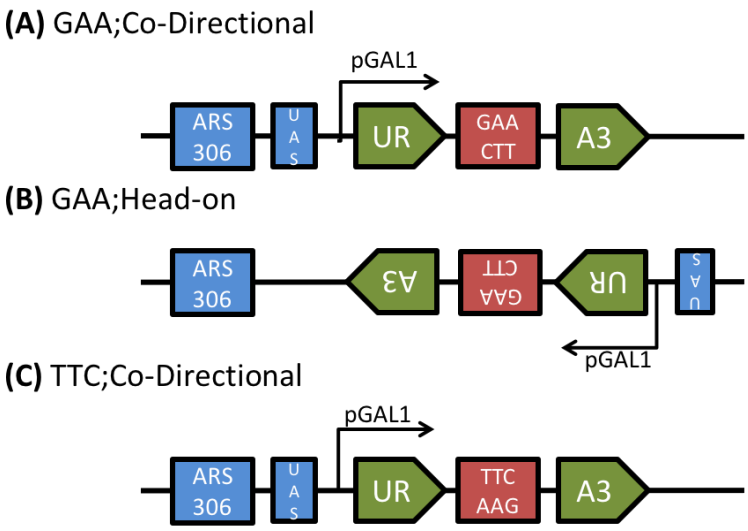
Yeast experimental system

In order to monitor GAA repeat expansion rate, we used genetic constructs that contained the *URA3* gene, which allowed for the selection of repeat expansion events (Figure 1A-C). The concepts of these cassettes was first described and implemented by Shishkin et al. 2009.

Cells with an active *URA3* gene are prototrophic for uracil and sensitive to 5-fluoroorotic acid (5-FOA). We inserted 100 GAA/TTC repeats into an artificial intron in the *URA3* gene in both orientations. When the repeat number expands beyond 100, the intron length increases beyond a threshold splicing length (~1000 bp), and the splicing efficiency of the *URA3* mRNA transcript decreases significantly, resulting in cells that are both Ura^- and $5\text{-}FOA^r$. Thus, Ura^- clones with expanded repeats can be selected for by plating on media containing 5-FOA. Our cassettes also contain an upstream yeast *GAL1* promoter, which allows us to induce a higher level of cassette transcription by growing the yeast on galactose and a lower level by growing on glucose.

Flanks of homology were used to integrate the cassettes into Chr III approximately 700 bp downstream from ARS 306, an early-firing origin of replication. Transformants were prototrophic for tryptophan and were selected for by plating on media lacking tryptophan. Successful transformation was confirmed with PCR analysis and sequencing. Our cassettes differ in three variables: identity of repeats on the lagging strand template during replication, identity of repeats on the sense strand during transcription, and direction of replication and transcription (Table 1). Strains will hereafter be referred to by the identity of the repeats on the

sense strand of transcription (GAA or TTC) and the relative direction of replication and transcription across the repeats (co-directional (CD) or head-on (HO)).



Cassette	Lagging strand template	mRNA transcript contains	Transcription/replication orientation	5-FOA concentration GLU/GAL
GAA;CD	GAA	GAA	Co-directional	0.110%/0.095%
GAA;HO	TTC	GAA	Head-on	0.110%/0.095%
TTC;CD	TTC	UUC	Co-directional	0.145%/0.125%

Figure 1. Selectable cassettes used to quantify repeat expansion. **A.** URA3 cassette containing 100 GAA repeats (on lagging strand template) in an artificial intron. **B.** URA3 cassette inverted with respect to ARS 306. TTC repeats located on lagging strand template and replication and transcription moving in head-on direction. **C.** URA3 cassette with GAA repeats inverted with respect to ARS 306. TTC repeats located on lagging strand template and replication and transcription moving co-directionally. **Table 1.** Characteristics of the three cassettes. Cassette names refer to the identity of the repeats sense strand of transcription and the relative direction of transcription and replication. 5-FOA concentration GLU/GAL refers to the concentration of 5-FOA required to select against cells without expanded repeats after growing on glucose or galactose for the non-selective phase.

The TTC;CD strain exhibited a higher 5-FOA resistance to start with compared to the strains with GAA;CD or GAA;HO cassettes. Previous work in the Mirkin lab theorized that this difference in resistance was due to the elevated degradation of mRNA transcripts containing

UUC repeats (Krasilnikova et al., 2007). To offset this effect, the 5-FOA concentration we used to select for expansions in the TTC cassette was 0.03% higher than the concentration used for the GAA cassette (Table 1).

Fluctuation assay

To measure the rate of GAA_{100}/TTC_{100} expansion per cell division, we used fluctuation tests. We grew strains of *S. cerevisiae* that contained our genetic cassettes overnight in liquid raffinose to remove all glucose and "de-repress" the *GAL1* promoter. Once washed in water, the yeast was plated on solid rich media containing glucose (YPD) for 40 hours or galactose (YPGal) for 48 hours. 12 independent colonies were then picked from each plate, dissolved in 200 ul sterile H_2O, and serially diluted. We plated appropriate dilutions on non-selective media (YPD) or selective media (5-FOA), followed by growth at 30°C. Following 3 days of growth, we counted the colonies on YPD and used the numbers to approximate the number of cells in the original colony. Following 3-4 days of growth, we counted the colonies on selective media and performed PCR analysis on an equal proportion of colonies from each plate (e.g. one-fourth) to confirm repeat expansion. On average, we analyzed a total of 96 colonies from each condition. Colonies with over 10 repeats added were considered expansion events (Figure 2). We used the number of colonies with expansions on selective media and the total number of colonies on rich media to calculate rate of expansion per cell division using the Ma-Sandri-Sarkar maximum likelihood method & FluCalc web software (Radchenko et al. 2018). Significance was also calculated using the rSalvador package in R, using 'Newton.LD.plating' and 'confint.LD.plating' commands (Zheng 2017).

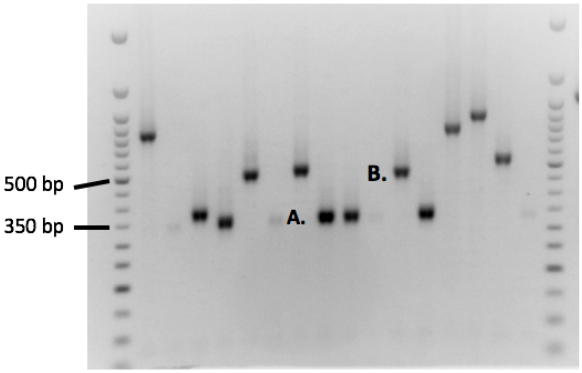
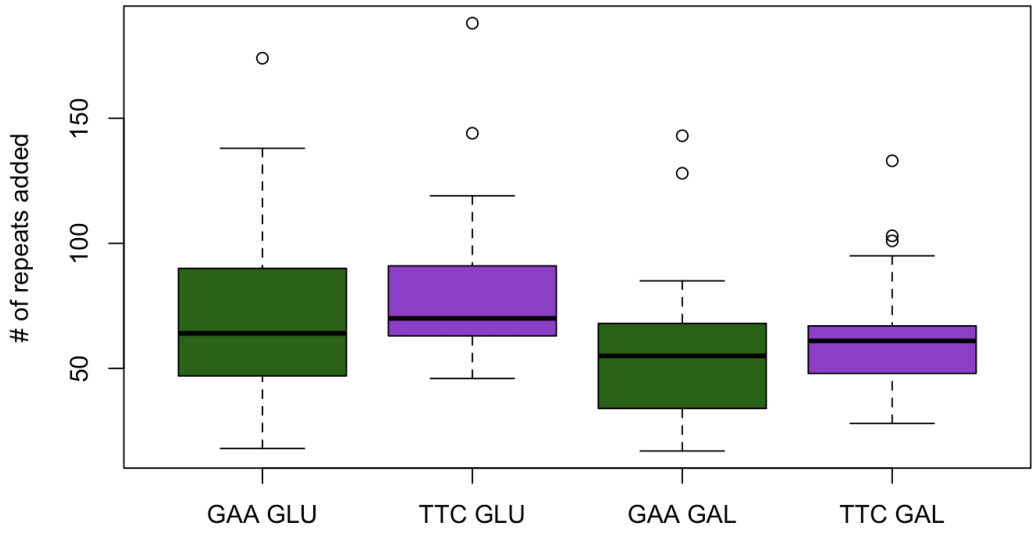


Figure 2. PCR fragments of GAA repeats in URA3 cassette run through a 1.5 % agarose gel with 50 bp ladder. **A.** 350 bp fragment – unexpanded repeat. **B.** 475 bp fragment – expanded repeat.

The expansion lengths observed in the 5-FOA resistant colonies were also collected to perform expansion length analysis. For this analysis, only non-clonal expansions were used to avoid skewing the data. If two expansions from the same 5-FOA resistant colony on the same plate have expansions of the same length, there is a strong possibility that they resulted from the same expansion event. Thus, if an expansion from one colony was within 10 bp of another expansion from the same colony, it was considered a clonal expansion and not included.

Results



Kolmogorov-Smirnov test			p =
GAA GLU	vs.	GAA GAL	0.214
TTC GLU	vs.	TTC GAL	0.127
GAA GLU	vs.	TTC GLU	0.273
GAA GAL	vs.	TTC GAL	0.421

Figure 3. Distribution of expansion lengths observed in fluctuation test performed on GAA;CD and TTC;CD cassettes. GLU represents low cassette transcription, and GAL represents high cassette transcription. **Table 2.** Results of the two-sided Kolmogorov-Smirnov test performed on comparisons between cassettes and between different transcription conditions.

There was no difference in expansion sizes between GAA and TTC cassettes

Strains containing the TTC;CD cassette exhibited a higher 5-FOA resistance than strains containing the GAA;CD or GAA;HO cassettes. This was observed earlier in Krasilnikova et al. 2007 and was theorized to be a result of UUC repeats in the URA3 mRNA transcript being a

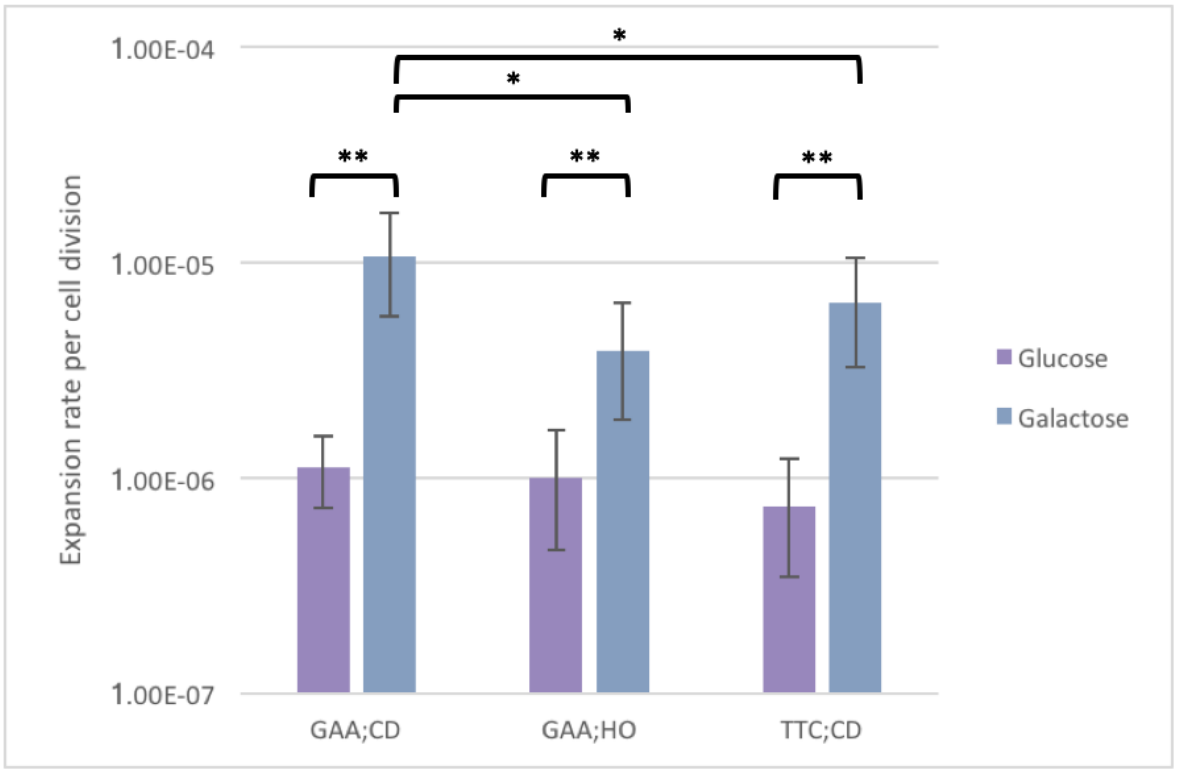
signal for degradation. As a result, we had to increase the 5-FOA concentration used to select for expanded repeats in strains containing the TTC;CD cassette. In order to ensure that the increased 5-FOA concentration was not selecting for longer repeat expansions and artificially altering our expansion rates, we performed two-sided Kolmogorov-Smirnov comparisons on the distributions of expansion lengths in our cassettes.

The TTC;CD cassette does appear to exhibit slightly longer expansion sizes than the GAA;CD cassette (Figure 3). However, there was no significant difference between the distributions. Thus, we felt comfortable comparing expansion rates between the GAA and TTC cassettes.

Transcription-mediated GAA repeat expansion is elevated when transcription and replication across the repeats are oriented co-directionally

We then performed fluctuation tests on strains containing our three experimental cassettes under conditions of low transcription (glucose) and conditions of high transcription (galactose) (Figure 4).

The three cassettes exhibited similar expansions rates under conditions of low transcription. Under conditions of high transcription, we observed small but significant decreases in expansion rate in GAA;HO and TTC;CD strains compared to the GAA;CD strain. This decrease in expansion rate in the GAA;HO cassette is consistent with the result observed in Shishkin et al. 2009. The feature that both the GAA;HO and TTC;CD cassettes share is that the lagging strand template contains TTC repeats instead of GAA repeats. Thus, we suggest that these decreases in expansion rate are due to some orientation-dependent mechanism that occurs during leading or lagging strand synthesis in conjunction with transcription.



Fold increase in expansion rate with galactose induction	
GAA;CD	9.58
GAA;HO	3.95
TTC;CD	8.81

Figure 4. GAA expansion rates of our experimental cassettes on glucose (conditions of low transcription) and galactose (conditions of high transcription). Error bars indicate 95% confidence intervals. * indicates non-overlapping 84% confidence intervals, and ** indicates non-overlapping 95% confidence intervals. **Table 2.** Fold increase in expansion rate upon galactose induction of transcription.

Transcription induced a significant increase in expansion rate in all three cassettes. We observed that transcription induction caused similar 8-10 fold increases in expansion rate in

both the GAA;CD and TTC;CD strains. This suggests that transcription-induced expansion is independent of whether the transcriptional sense strand contains GAA or TTC repeats.

We also observed that inducing transcription in our co-directional cassettes (GAA;CD & TTC;CD) caused an 8-10 fold increase in expansion rate, while inducing transcription in our head-on cassette (GAA;HO) caused a lesser 3.95-fold increase in expansion rate (Table 2). Given that the effect of transcription on expansion rate was stronger in the co-directional cassettes than in the head-on cassette, we concluded that some aspect of orienting transcription and replication of the repeats co-directionally promotes GAA repeat expansion. Consequently, we can also rule out head-on collisions between transcription and replication as a major contributor to GAA repeat expansion in our system.

R-loop mediated repeat expansion is orientation-independent

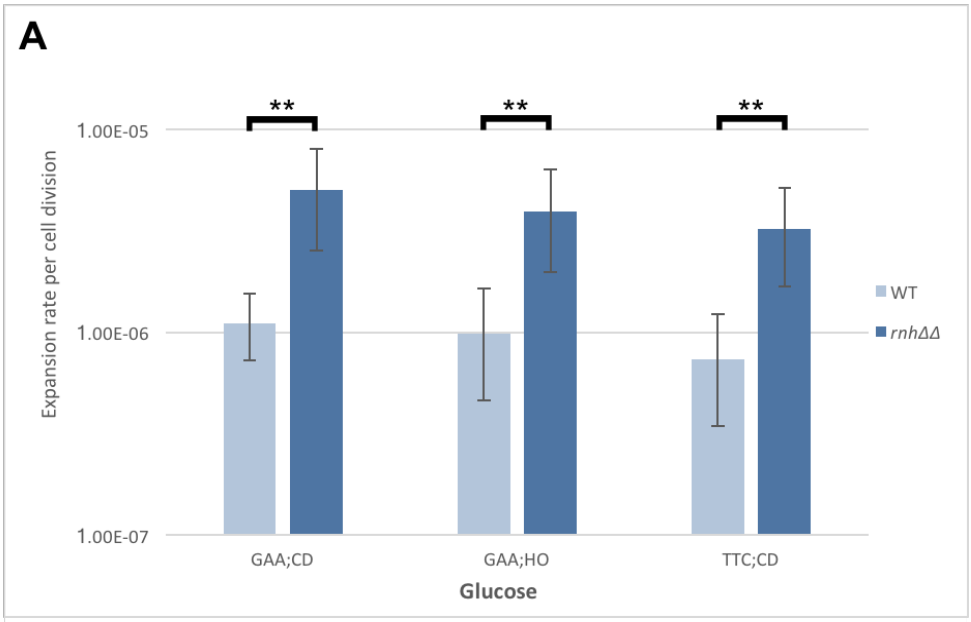
We also performed fluctuation tests on *rnh1Δ;rnh201Δ* strains containing one of our three experimental cassettes under conditions of low transcription (glucose) and high transcription (galactose). We then compared these expansion rates to those of the WT (Figure 5).

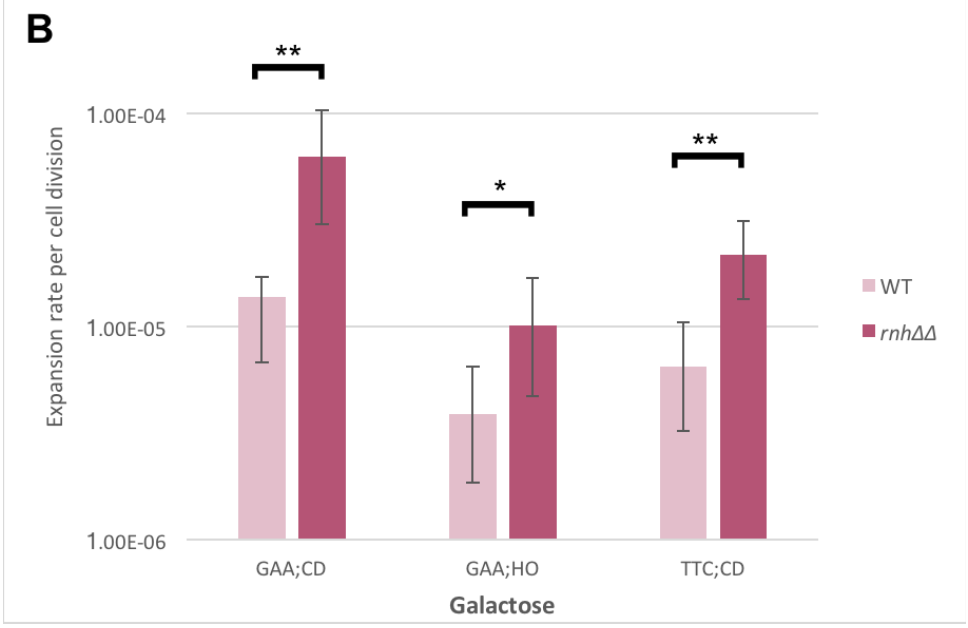
Under conditions of low transcription, knocking out RNase H enzymes induced a very similar 4-5 fold increase in expansion rate in all three cassettes. Under conditions of high transcription, knocking out RNase H enzymes also induced a significant increase in expansion rate in all three cassettes, albeit of varying scales.

We expected the TTC;CD cassette with its pyrimidine-rich sense strand to be less conducive to R-loop formation and therefore, we predicted that the expansion rate would be unaffected by knocking out RNase H enzymes. However, we did observe a significant increase in expansion rate of the TTC;CD cassette upon knocking out the RNase H enzymes. We should note that the fold increase over WT in expansion rate caused by the RNase H double knockout grown on galactose was smaller in the TTC;CD cassette (3.34) than in the GAA;CD cassette

(5.91) (Table 3). Still, we can conclude that R-loop-mediated expansions occur regardless of whether the sense strand of the repeats is homopurine or homopyrimidine.

We also hypothesized that the RNase H knockout could have a stronger effect on expansion rate of the GAA;HO cassette because it has been observed that head-on collisions between transcription and replication can induce R-loop formation (Hamperl et al. 2017). However, we observed that the RNase H knockout had the least effect on expansion rate in the GAA;HO cassette (2.58-fold increase) (Table 3). Thus, we concluded that head-on collisions between transcription and replication are not necessary to induce the R-loops that result in R-loop-mediated repeat expansion.





Fold increase in expansion rate with *rnh1;rnh201* double knockout					
Glucose			Galactose		
GAA;CD	GAA;HO	TTC;CD	GAA;CD	GAA;HO	TTC;CD
4.51	4.00	4.40	4.55	2.58	3.34

Figure 5. GAA expansion rates of our experimental cassettes in WT and *rnhΔΔ* strains. **A.** Fluctuation test on glucose (conditions of low transcription). ** indicates non-overlapping 95% confidence intervals. * indicates non-overlapping 84% confidence intervals. **B.** Fluctuation test on galactose (conditions of high transcription) **Table 3.** Fold increase in expansion rate upon double knockout of *rnh1* and *rnh201* compared to WT.

Discussion

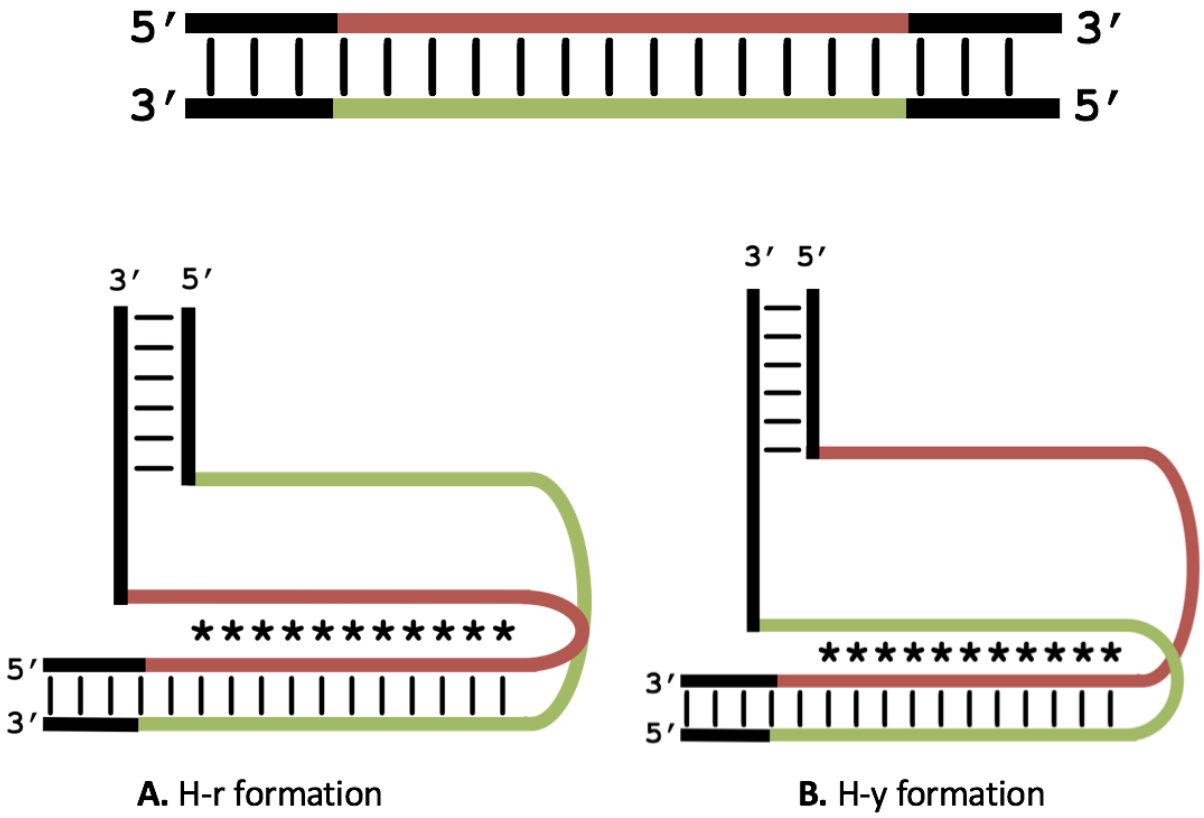


Figure 6. Two possible conformations of H-DNA in tracts of homopurine/homopyrimidine mirror repeats. Red represents homopurine segments, green represents homopyrimidine segments. | represents Watson-Crick pairing. * represents Hoogsteen or reverse-Hoogsteen base pairing. **A.** H-r formation where the third strand in the triplex is homopurine and antiparallel to the complementary strand. **B.** H-y formation where the third strand in the triplex is homopyrimidine and parallel to the complementary strand.

From what we have observed about the orientation-dependence of repeat expansion, we have proposed a mechanism for how R-loop-mediated expansions occur based on the propensity of GAA repeats to form H-DNA.

H-DNA: A non-B DNA conformation

H-DNA is a secondary DNA structure that forms in tracts of homopurine/ homopyrimidine (PuPy) mirror repeats (Frank-Kamenetskii & Mirkin 1995). When the sequence experiences negative supercoiling, one strand can fold back into the major groove of the DNA duplex,

consequently forming a triplex helix adjacent to a single strand of DNA (Figure 6). The third strand of the triplex is bonded to the homopurine strand by Hoogsteen hydrogen bonds, which are an alternative way for nitrogenous bases to pair with each other. Purines are capable of forming Watson-Crick and Hoogsteen hydrogen bonds simultaneously (Frank-Kamenetskii & Mirkin 1995). Disruptions to the mirror symmetry of the repeat will significantly decrease the favorability of H-DNA formation (Mirkin et al. 1987).

H-DNA can form in two conformations: H-r (Figure 6A) or H-y (Figure 6B). In the H-r conformation, the third strand in the triplex is homopurine and binds antiparallel to the other homopurine strand through reverse-Hoogsteen hydrogen bonds (Du & Zhou 2013). This triplex is composed of C–G*G and T–A*A bonds (Bacolla et al. 2015). (– represents Watson-Crick base pairing, * represents Hoogsteen or reverse-Hoogsteen bonds.)

In the H-y formation, the third strand in the triplex is homopyrimidine and binds parallel to the homopurine strand through Hoogsteen bonds (Du & Zhou 2013). This triplex is composed of C–G*C^+ and T–A*T bonds (Bacolla et al. 2015). Because the Hoogsteen bond between the guanine and cytosine requires the cytosine to be protonated at N3, the formation of this triplex is favored at lower pH levels (Mirkin & Frank-Kamenetskii 1994).

There is evidence that H-DNA plays a biological role in the cell. In humans, potential H-DNA forming sequences were observed to occur every 50,000 bp (Schroth & Ho 1995). These PuPy sequences have been observed particularly in promoter sequences (Wells et al. 1988, Lu & Ferl 1993). In addition, H-DNA forming sequences have been implicated in the transcription regulation of various genes, including the human proto-oncogene *c-myc* (Brahmachari et al. 1996, Kinniburgh et al. 1994).

However, H-DNA can also be deleterious. H-DNA analog structures blocked transcription elongation *in vitro* (Pandey et al. 2015). Furthermore, H-DNA forming sequences from the human genome were shown to induce gross deletions in *E. coli* plasmids (Bacolla et al.

2004) and DNA double strand breaks and mutations in mammalian cells (Wang & Vasquez 2004).

Importantly, GAA repeats are PuPy mirror repeats and thus have the propensity to form H-DNA. GAA repeats have been shown to form H-DNA *in vitro* at physiological pH (Mariappan et al. 1999). Furthermore, atomic force microscopy has been used to visualize the formation of H-DNA in GAA repeats in plasmids (Potaman et al. 2004). The tendency of GAA to form H-DNA could also play a role in the etiology of FRDA. The formation of a triplex in GAA repeats has been shown to exclude nucleosome assembly *in vitro*, which could contribute to transcriptional silencing of the *FXN* gene *in vivo* (Ruan & Wang 2008). Furthermore, the addition of an oligodeoxyribonucleotide designed to disrupt triplex formation was shown to increase transcription through GAA repeats *in vitro* (Grabczyk & Usdin 2000).

Proposed H-loop mechanism for repeat expansion

Based on our data, we propose that R-loop-mediated repeat expansion is caused by the transcribed GAA repeats in our experimental system forming an H-loop: a structure that combines an R-loop with H-DNA (Figure 7). In our model, RNA Polymerase II transcribes the repeats and induces negative supercoiling behind it, which makes triplex formation and R-loop formation more favorable (Masseé & Drolet 1999, Mirkin & Frank-Kamenetskii 1994). Consequently, H-DNA forms, and the mRNA transcript of the repeats stays bound to the ssDNA portion, stabilizing the entire structure—an H-loop. This structure would form an effective block to replication fork progression, which could then cause repeat instability.

There is evidence that it is possible for such a structure to form. The addition of an oligonucleotide complementary to the single-stranded portion of H-DNA *in vitro* was shown to stabilize the formation of an H-y triplex up to a pH of 7 (Belotserkovskii et al. 1992). In addition, electron microscopy of a transcribed PuPy sequence *in vitro* suggests that the mRNA transcript is capable of binding to the single-stranded portion of H-DNA (Grabczyk & Fishman 1995). Lastly, Grabczyk & Usdin 2000 showed that an H-y triplex formed from GAA/TTC repeats will

bind to an oligodeoxyribonucleotide that is complementary to the single-stranded portion of the triplex.

H-loops would form readily in both GAA and TTC repeats, explaining why we observed R-loop-mediated repeat expansion regardless of repeat orientation. GAA repeats can form H-DNA in both H-r and H-y configurations (Vetcher et al. 2002, Potaman et al. 2004, Grabczyk & Usdin 2000). We propose that in all three of our experimental cassettes, an H-loop forms in an orientation such that the transcriptional template strand is single stranded and can then bind to the mRNA strand. Thus, the repeats in the GAA;HO and GAA;CD cassettes form H-DNA in an H-r configuration (Figure 7A and 7B), and the repeats in the TTC;CD cassette form H-DNA in an H-y configuration (Figure 7C).

The RNA-DNA hybrid formed in the TTC;CD cassette (GAA-RNA, TTC-DNA) is theoretically less stable than the RNA-DNA hybrid formed in the GAA;CD and GAA;HO cassettes (UUC-RNA, GAA-DNA); however, because of the formation of the triplex, the RNA strand does not have to compete with the complementary DNA strand to form the R-loop. Thus, the H-loop would be able to form in both the GAA and TTC cassettes. This is consistent with our observation that R-loop-mediated repeat expansion is observed in GAA;CD and TTC;CD cassettes (Figure 5B).

Our H-loop model also explains why the expansion rates of the GAA;HO cassette were different from expansion rates of the GAA;CD and TTC;CD cassettes. In WT strains, we observed that transcription induction increased expansion rate to a smaller degree in the GAA;HO cassette compared to the expansion rates in the GAA;CD and TTC;CD cassettes (Figure 4). In addition, we observed that the RNase H double knockout had a smaller (though still significant) effect on the expansion rate of the GAA;HO cassette compared to the expansion rates of the GAA;CD and TTC;CD cassettes (Figure 5).

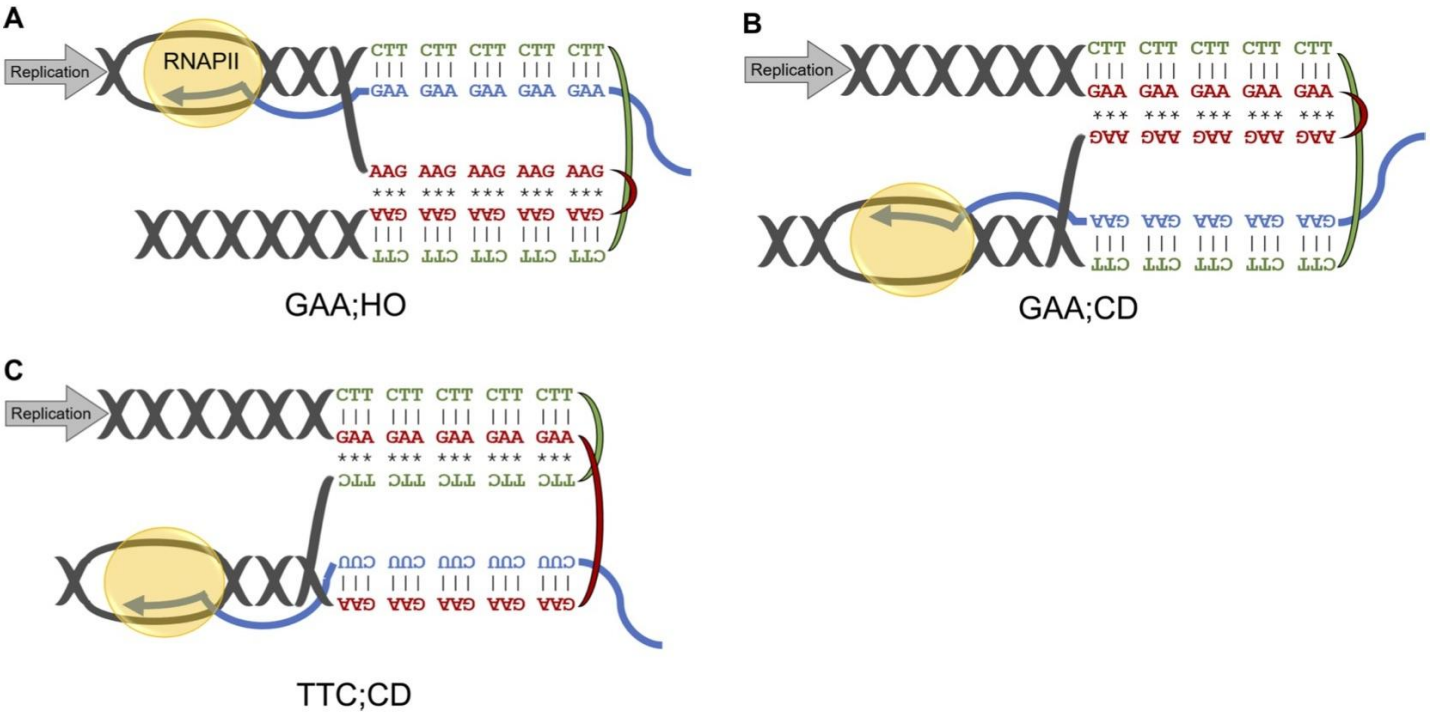


Figure 7. Proposed H-loop formation in the 3 different cassettes. Red represents homopurine segments, green represents homopyrimidine segments, blue represents mRNA. | represents Watson-Crick pairing. * represents Hoogsteen or reverse-Hoogsteen base pairing. **A.** GAA;HO cassette. H-DNA is in H-r formation. DNA replication encounters the R-loop first. **B.** GAA;CD cassette. H-DNA is in H-r formation. DNA replication encounters the triplex first. **C.** TTC;CD cassette. H-DNA is in H-y formation. DNA replication encounters the triplex first.

These observations are consistent with our model. In our experimental system, the cassettes are positioned such that replication will proceed from the upstream direction. In the GAA;HO cassette, replication would encounter the RNA-DNA hybrid first (Figure 7A), while in the GAA;CD and TTC;CD cassettes, replication would encounter the triplex first (Figure 7B and 7C). Replicative helicases are fully capable of unwinding an RNA-DNA duplex, when translocating along the DNA strand, as would be the case in an H-loop (Shin & Kelman 2006). With the R-loop removed first, the rest of the triplex would then be less stable and be a less effective block to the replication fork. On the other hand, triplexes have been observed to impede unwinding by a eukaryotic helicase (Peleg et al 1995). Thus, the H-loop formed in the

GAA;HO cassette would theoretically be a less effective block to replication than the H-loops formed in the other cassettes.

As to the cause of the repeat expansion, we propose that the H-loop causes the collapse of the replication fork, resulting in a one-ended double strand break (DSB) that is repaired erroneously into an expansion by break-induced replication (BIR). R-loop-induced expansion is also Rad52- and Pol32-dependent in *S. cerevisiae* (Neil et al. 2018). Pol32, a subunit of Pol δ, and Rad52, a mediator of strand exchange, are both required for most cases of BIR (Anand et al. 2013). In BIR, the 3' end of a one-ended DSB invades the homologous chromosome and uses it as a template to synthesize the rest of the chromosome (Anand et al. 2013). If the strand invades the repetitive region of the homologous chromosome "out-of-register" (e.g. it invades the region upstream of where it should), then BIR would result in a repeat expansion. An out-of-register invasion during BIR has also been has also been implicated in expansions of CAG repeats in yeast (Kim et al. 2016).

A future experiment that would provide additional evidence for the H-loop model would be to create a TTC;HO cassette by inverting the TTC;CD cassette with respect to ARS 306. Based on our H-loop model, we predict that this cassette will exhibit the smallest effect of transcription induction and the smallest effect of the RNase H double knockout due to its poly-pyrimidine sense strand and co-directional orientation.

Furthermore, a worthy avenue of investigation would be the role that leading or lagging strand synthesis (in combination with transcription) may play in repeat expansion. These processes were implicated by our observation that both the GAA;HO and TTC;CD cassettes (TTC repeats on lagging strand template) exhibited significant decreases in expansion rate compared to the GAA;CD (GAA repeats on lagging strand template) cassette under conditions of high transcription (Figure 3). Our data seem to suggest that some aspect of locating the GAA repeats on the lagging strand template or the TTC repeats on the leading strand template when in combination with high-levels of transcription can induce repeat expansion.

Shah et al. 2014 showed that transcription at or around GAA repeats will decrease the nucleosome density at the repeats, which could increase the probability of template switching. When template switching occurs during replication, a nascent strand can switch to using the other nascent strand as a template in order to bypass a lesion. If the template switch occurs out-of-register, this could result in an expansion. Thus, if template switching is dependent on the identity of the repeats on the lagging strand template, then we would observe lower expansion rates in the GAA;HO and TTC;CD cassettes compared to the GAA;CD cassette but only under conditions of high transcription (Figure 4). The template switching will be further explored in Chapter 2.

In conclusion, we have shown that transcription-induced and R-loop-mediated GAA repeat expansions occur independently of the relative directions of transcription and replication, as well as the composition of the sense strand of transcription. We have also shown that transcription induces a stronger increase in expansion rate when it is oriented co-directionally with replication across the repeats. From these conclusions, we propose that transcribed GAA repeats can block replication and induce repeat expansion by forming H-loops in an orientation-independent manner.

Citations

Anand, R. P., Lovett, S. T., & Haber, J. E. (2013). Break-Induced DNA Replication. *Cold Spring Harb. Perspect. Biol., 5*(12). doi:10.1101/cshperspect.a010397

Bacolla, A., Jaworski, A., Larson, J. E., Jakupciak, J. P., Chuzhanova, N., Abeysinghe, S. S., … Wells, R. D. (2004). Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc. Natl. Acad. Sci. U.S.A., 101*(39), 14162-14167. doi:10.1073/pnas.0405974101

Bacolla, A., Wang, G., & Vasquez, K. M. (2015). New Perspectives on DNA and RNA Triplexes As Effectors of Biological Activity. *PLoS Genet., 11*(12). doi:10.1371/journal.pgen.1005696

Belotserkovskii, B. P., Krasilnikova, M. M., Veselkov, A. G., & Frank-Kamenetskii, M. D. (1992). Kinetic trapping of H-DNA by oligonucleotide binding. *Nucleic Acids Res.,20*(8), 1903-1908. doi:10.1093/nar/20.8.1903

Belotserkovskii, B. P., Neil, A. J., Saleh, S. S., Shin, J. H., Mirkin, S. M., & Hanawalt, P. C. (2012). Transcription blockage by homopurine DNA sequences: role of sequence composition and single-strand breaks. *Nucleic Acids Res., 41*(3), 1817-1828. doi:10.1093/nar/gks1333

Brahmachari, S. K., Sarkar, P. S., Raghavan, S., Narayan, M., & Maiti, A. K. (1997). Polypurine/polypyrimidine sequences as cis-acting transcriptional regulators. *Gene, 190*(1), 17-26. doi:10.1016/s0378-1119(97)00034-6

Campuzano, V., Montermini, L., Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., . . . Pandolfo, M. (1996). Friedreich's Ataxia: Autosomal Recessive Disease Caused by an Intronic GAA Triplet Repeat Expansion. *Science, 271*(5254), 1423-1427. doi:10.1126/science.271.5254.1423

Cossée, M., Schmitt, M., Campuzano, V., Reutenauer, L., Moutou, C., Mandel, J., & Koenig, M. (1997). Evolution of the Friedreichs ataxia trinucleotide repeat expansion: Founder effect and premutations. *Proc. Natl. Acad. Sci. U.S.A., 94*(14), 7452-7457. doi:10.1073/pnas.94.14.7452

DeJesus-Hernandez, M., Mackenzie, I., Boeve, B., Boxer, A., Baker, M., Rutherford, N., . . . Rademakers, R. (2011). Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron, 72*(2), 245-256. doi:10.1016/j.neuron.2011.09.011

Du, Y., & Zhou, X. (2013). Targeting Non-B-Form DNA in Living Cells. *Chem. Rec., 13*(4), 371-384. doi:10.1002/tcr.201300005

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet., 5*(6), 435-445. doi:10.1038/nrg1348

Filla, A., Michele, G. D., Cavalcanti, F., Pianese, L., Monticelli, A., Campanella, G., & Cocozza, S. (1996). The relationship between trinucleotide (GAA) repeat length and clinical features in Friedreich ataxia. *Am. J. Hum. Genet.,* 554-560.

Frank-Kamenetskii, M. D., & Mirkin, S. M. (1995). Triplex DNA Structures. *Ann. Rev. Biochem., 64*(1), 65-95. doi:10.1146/annurev.bi.64.070195.000433

Gan, W., Guan, Z., Liu, J., Gui, T., Shen, K., Manley, J. L., & Li, X. (2011). R-loop-mediated genomic instability is caused by impairment of replication fork progression. *Genes Dev., 25*(19), 2041-2056. doi:10.1101/gad.17010011

Gerhardt, J., Bhalla, A., Butler, J., Puckett, J., Dervan, P., Rosenwaks, Z., & Napierala, M. (2016). Stalled DNA Replication Forks at the Endogenous GAA Repeats Drive Repeat

Expansion in Friedreich's Ataxia Cells. *Cell Rep., 16*(5), 1218-1227. doi:10.1016/j.celrep.2016.06.075

Grabczyk, E., & Fishman, M. C. (1995). A Long Purine-Pyrimidine Homopolymer Acts as a Transcriptional Diode. *J. Biol. Chem., 270*(4), 1791-1797. doi:10.1074/jbc.270.4.1791

Grabczyk, E., & Usdin, K. (2000). Alleviating transcript insufficiency caused by Friedreich's ataxia triplet repeats. *Nucleic Acids Res., 28*(24), 4930-4937. doi:10.1093/nar/28.24.4930

Groh, M., Lufino, M. M. P., Wade-Martins, R., & Gromak, N. (2014). R-loops Associated with Triplet Repeat Expansions Promote Gene Silencing in Friedreich's Ataxia and Fragile X Syndrome. *PLoS Genet.*, *10*(5), e1004318. http://doi.org/10.1371/journal.pgen.1004318

Groh, M., & Gromak, N. (2014). Out of Balance: R-loops in Human Disease. *PLoS Genet., 10*(9). doi:10.1371/journal.pgen.1004630

Gyi, J. I., Conn, G. L., Lane, A. N., & Brown, T. (1996). Comparison of the Thermodynamic Stabilities and Solution Conformations of DNA·RNA Hybrids Containing Purine-Rich and Pyrimidine-Rich Strands with DNA and RNA Duplexes. *Biochemistry, 35*(38), 12538-12548. doi:10.1021/bi960948z

Hamperl, S., Bocek, M. J., Saldivar, J. C., Swigut, T., & Cimprich, K. A. (2017). Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell, 170*(4). doi:10.1016/j.cell.2017.07.043

Kim, J. C., Harris, S. T., Dinter, T., Shah, K. A., & Mirkin, S. M. (2016). The role of break-induced replication in large-scale expansions of $(CAG)_n/(CTG)_n$ repeats. *Nat. Struct. Mol. Biol., 24*(1), 55-60. doi:10.1038/nsmb.3334

Kim, H., Narayanan, V., Mieczkowski, P. A., Petes, T. D., Krasilnikova, M. M., Mirkin, S. M., & Lobachev, K. S. (2008). Chromosome fragility at GAA tracts in yeast depends on repeat orientation and requires mismatch repair. *The EMBO Journal, 27*(21), 2896-2906. doi:10.1038/emboj.2008.205

Kinniburgh, A. J., Firulli, A. B., & Kolluri, R. (1994). DNA triplexes and regulation of the c-myc gene. *Gene,149*(1), 93-100. doi:10.1016/0378-1119(94)90416-2

Kovtun, I. V., & McMurray, C. T. (2008). Features of trinucleotide repeat instability in vivo. *Cell Res.,18*(1), 198-213. doi:10.1038/cr.2008.5

Krasilnikova, M. M., Kireeva, M. L., Petrovic, V., Knijnikova, N., Kashlev, M., & Mirkin, S. M. (2007). Effects of Friedreich's ataxia $(GAA)_n$•$(TTC)_n$ repeats on RNA synthesis and stability. *Nucleic Acids Res., 35*(4), 1075-1084. doi:10.1093/nar/gkl1140

Lander E., Linton L., Birren B., Nusbaum C., Zody M., Baldwin J, ... International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature,* 409(6822):860–921.

Li, Y., Lu, Y., Polak, U., Lin, K., Shen, J., Farmer, J., Seyer, L., Bhalla, A. D., Rozwadowska, N., Lynch, D., Sergesketter Butler, J., Napierala, M. (2015). Expanded GAA repeats impede transcription elongation through the *FXN* gene and induce transcriptional silencing that is restricted to the *FXN* locus. *Hum. Mol. Genet.*, *24*(24), 6932–6943. http://doi.org/10.1093/hmg/ddv397

Lu, G., & Ferl, R. J. (1993). Homopurine/homopyrimidine sequences as potential regulatory elements in eukaryotic cells. *Int. J. Biochem., 25*(11), 1529-1537. doi:10.1016/0020-711x(93)90508-c

Lupoli, F., Vannocci, T., Longo, G., Niccolai, N., & Pastore, A. (2017). The role of oxidative stress in Friedreich's ataxia. *FEBS Letters*. doi:10.1002/1873-3468.12928

Mariappan, S., Catasti, P., Silks, L. A., Bradbury, E., & 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*(5), 2035-2052. doi:10.1006/jmbi.1998.2435

Massé, E., & Drolet, M. (1999). *Escherichia coli* DNA Topoisomerase I Inhibits R-loop Formation by Relaxing Transcription-induced Negative Supercoiling. *J. Biol. Chem., 274*(23), 16659-16664. doi:10.1074/jbc.274.23.16659

Mirkin, S. M., Lyamichev, V. I., Drushlyak, K. N., Dobrynin, V. N., Filippov, S. A., & Frank-Kamenetskii, M. D. (1987). DNA H form requires a homopurine–homopyrimidine mirror repeat. *Nature, 330*(6147), 495-497. doi:10.1038/330495a0

Mirkin, S. M., & Frank-Kamenetskii, M. D. (1994). H-DNA and Related Structures. *Ann. Rev. Biophys. Biomol. Struct., 23*(1), 541-576. doi:10.1146/annurev.bb.23.060194.002545

Mirkin, S. M. (2007). Expandable DNA repeats and human disease. *Nature, 447*(7147), 932-940. doi:10.1038/nature05977

Montermini, L., Richeter, A., Morgan, K., Justice, C. M., Julien, D., Castellotti, B., . . . Pandolfo, M. (1997). Phenotypic variability in friedreich ataxia: Role of the associated GAA triplet repeat expansion. *Ann. of Neurol., 41*(5), 675-682. doi:10.1002/ana.410410518

Neil, A. J., Liang, M. U., Khristich, A. N., Shah, K. A., & Mirkin, S. M. (2018). RNA–DNA hybrids promote the expansion of Friedreich's ataxia $(GAA)_n$ repeats via break-induced replication. *Nucleic Acids Res*., doi:10.1093/nar/gky099

Pandey, S., Ogloblina, A. M., Belotserkovskii, B. P., Dolinnaya, N. G., Yakubovskaya, M. G., Mirkin, S. M., & Hanawalt, P. C. (2015). Transcription blockage by stable H-DNA analogs in vitro. *Nucleic Acids Res., 43*(14), 6994-7004. doi:10.1093/nar/gkv622

Pandolfo, M. (2001). Molecular basis of Friedreich ataxia. *Mov. Disord.,16*(5), 815-821. doi:10.1002/mds.1162

Peleg, M., Kopel, V., Borowiec, J. A., & Manor, H. (1995). Formation of DNA triple helices inhibits DNA unwinding by the SV40 large T-antigen helicase. *Nucleic Acids Res., 23*(8), 1292-1299. doi:10.1093/nar/23.8.1292

Potaman, V. N., Oussatcheva, Y. L., Lyubchenko, L. S., Shlyakhtenko, S. I., Bidichandani, T. Ashizawa, Sinden, R. R. (2004). Length-dependent structure formation in Friedreich ataxia (GAA)n·(TTC)n repeats at neutral pH. *Nucleic Acids Res., 32*(3), 1224-1231. doi:10.1093/nar/gkh274

Prado, F., & Aguilera, A. (2005). Impairment of replication fork progression mediates RNA pol II transcription-associated recombination. *The EMBO Journal, 24*(6), 1267-1276. doi:10.1038/sj.emboj.7600602

Radchenko, E. A., Mcginty, R. J., Aksenova, A. Y., Neil, A. J., & Mirkin, S. M. (2018). Quantitative Analysis of the Rates for Repeat-Mediated Genome Instability in a Yeast Experimental System. *Methods Mol. Biol.,* 421-438. doi:10.1007/978-1-4939-7306-4_29

Rindler, P. M., & Bidichandani, S. I. (2010). Role of transcript and interplay between transcription and replication in triplet-repeat instability in mammalian cells. *Nucleic Acids Res., 39*(2), 526-535. doi:10.1093/nar/gkq788

Rocha, E. P., & Danchin, A. (2003). Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res., 31*(22), 6570-6577. doi:10.1093/nar/gkg859

Ruan, H., & Wang, Y. (2008). Friedreich's Ataxia GAA·TTC Duplex and GAA·GAA·TTC Triplex Structures Exclude Nucleosome Assembly. *J. Mol. Biol., 383*(2), 292-300. doi:10.1016/j.jmb.2008.08.053

Schroth, G. P., & Ho, P. S. (1995). Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic Acids Res., 23*(11), 1977-1983. doi:10.1093/nar/23.11.1977

Shah, K. A., McGinty, R. J., Egorova, V. I., & Mirkin, S. M. (2014). Coupling transcriptional state to large-scale repeat expansions in yeast. *Cell Rep.*, *9*(5), 1594–1602. http://doi.org/10.1016/j.celrep.2014.10.048

Sharma, R., Bhatti S., Gomez M., Clark R.M., Murray C., Ashizawa T., Bidichandani, S.I. (2002). The GAA triplet-repeat sequence in Friedreich ataxia shows a high level of

somatic instability in vivo, with a significant predilection for large contractions. *Hum. Mol. Genet., 11*(18), 2175-2187. doi:10.1093/hmg/11.18.2175

Shin, J., & Kelman, Z. (2006). The Replicative Helicases of Bacteria, Archaea, and Eukarya Can Unwind RNA-DNA Hybrid Substrates. *J. Biol. Chem., 281*(37), 26914-26921. doi:10.1074/jbc.m605518200

Shishkin, A. A., Voineagu, I., Matera, R., Cherng, N., Chernet, B. T., Krasilnikova, M. M., Narayanan, V., Lobachev, K., Mirkin, S. M. (2009). Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. *Mol. Cell*, *35*(1), 82–92. http://doi.org/10.1016/j.molcel.2009.06.017

Su, X. A., & Freudenreich, C. H. (2017). Cytosine deamination and base excision repair cause R-loop–induced CAG repeat fragility and instability in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci.,114*(40). doi:10.1073/pnas.1711283114

Subramanian, S., Mishra, R. K., & Singh, L. (2003). Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol., 4*(2), R13. doi:10.1186/gb-2003-4-2-r13

Sutherland, G. R., & Richards, R. I. (1995). Simple tandem DNA repeats and human genetic disease. *Proc. Natl. Acad. Sci. U.S.A., 92*(9), 3636-3641. doi:10.1073/pnas.92.9.3636

Takeuchi, Y., Horiuchi, T., & Kobayashi, T. (2003). Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Genes Dev., 17*(12), 1497-1506. doi:10.1101/gad.1085403

Vetcher, A. A., Napierala, M., Iyer, R. R., Chastain, P. D., Griffith, J. D., & 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*(42), 39217-39227. doi:10.1074/jbc.m205209200

Wang, G., & Vasquez, K. M. (2004). Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A., 101*(37), 13448-13453. doi:10.1073/pnas.0405116101

Wells, R. D., Collier, D. A., Hanvey, J. C., Shimizu, M., & Wohlrab, F. (1988). The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences. *The FASEB Journal, 2*(14), 2939-2949. doi:10.1096/fasebj.2.14.3053307

Zhao, X., & Usdin, K. (2015). The Repeat Expansion Diseases: The dark side of DNA repair. *DNA Repair, 32*, 96-105. doi:10.1016/j.dnarep.2015.04.019

Zheng, Q. (2017). RSalvador: An R Package for the Fluctuation Experiment. *G3(Bethesda), 7*(12), 3849-3856. doi:10.1534/g3.117.300120

Chapter 2 - Okazaki flap processing and GAA repeat expansion

Abstract

During the synthesis of an Okazaki fragment, Pol δ displaces the 5' end of the preceding fragment, creating a 5' flap that is processed by the flap endonuclease Rad27 in *S. cerevisiae* (FEN1 in humans). Rad27 is integral to genome stability, and mutations in *rad27* can induce up to a 100-fold increase in GAA repeat expansion rate (Tsutakawa et al. 2017). The mechanism of these flap-associated expansions is still unknown. We wanted to test the hypothesis that these expansions were caused by the template switch pathway of post-replicative repair (PRR). To do so, we looked at the role of the Rad5 protein, a ubiquitin-ligase necessary to initiate the process of template switch. We also used Rad27 hypomorph strains 2E and 4A, described by Tsutakawa et al. 2017. To monitor expansion rate, we utilized two selectable cassettes that each contain 100 GAA repeats, one detecting expansions >10 repeats (all-scale), and another one detecting expansions of >60 repeats (large-scale). We performed fluctuation tests on WT strains, *rad5Δ* single mutants*, rad27* single mutants, and *rad27;rad5Δ* double mutants containing one of our selectable cassettes. We observed that Rad5 knockout partially rescued the elevated expansion rate phenotype in the *rad27* mutants when looking at all-scale expansions but not when looking at large-scale expansions alone. Thus, we concluded that small-scale flap-associated expansions are Rad5-dependent. To explain this, we proposed template switch and fork regression mechanisms for repeat expansion. We also concluded that large-scale flap-associated expansions are Rad5-independent. To explain this, we proposed a model of repeat expansion based on break-induced replication (BIR).

Introduction

Okazaki flap processing and Rad27

All DNA polymerases replicate from 5' to 3'— they catalyze the attachment of the 3' hydroxyl of a growing DNA strand to the 5' phosphate of a deoxyribonucleotide. Due to the antiparallel nature of the DNA duplex, one of the strands — the lagging strand — must be replicated in a discontinuous manner. The lagging strand is synthesized in segments called Okazaki fragments, which are ~165 bp long on average (Smith & Whitehouse 2012). These separate fragments must then be processed into a cohesive strand of DNA.

This process begins when DNA polymerase, synthesizing a downstream Okazaki fragment, displaces the 5' end of the upstream Okazaki fragment. The displaced end will hang off the DNA as a flap until it is processed. Once the flap is removed, there remains a single stranded break in the DNA backbone that DNA ligase I can then join together.

Flap processing has been observed to occur through three different mechanisms. The main pathway involves an enzyme named FEN1 in humans (Rad27 is the *S. cerevisiae* analog) (Zheng & Shen 2011). FEN1 processes Okazaki fragments during lagging strand synthesis by cleaving the 5' end of the flap at its junction with the double stranded part of the downstream Okazaki fragment (Chapados et al. 2004). FEN1 has several positively charged residues in a helical gateway region that allow it to thread the negatively charged phosphate backbone of the DNA flap through itself to cleave at the proper place (Tsutakawa et al. 2017).

FEN1 is integral to maintaining genome stability. One study observed a 63-fold increase in mutations in *rad27* mutants compared to WT (Tishkoff et al. 1997). Furthermore, FEN1 contributes to telomere maintenance in humans and yeast (Saharia et al. 2008, Parenteau & Wellinger 2002) and plays a role in long-patch base excision repair, which can repair damaged DNA that cannot be processed by regular base excision repair (Gary et al. 1999). Lastly, mutations in *FEN1* as well as *FEN1* overexpression have been observed in various types of cancer (Zheng et al. 2010).

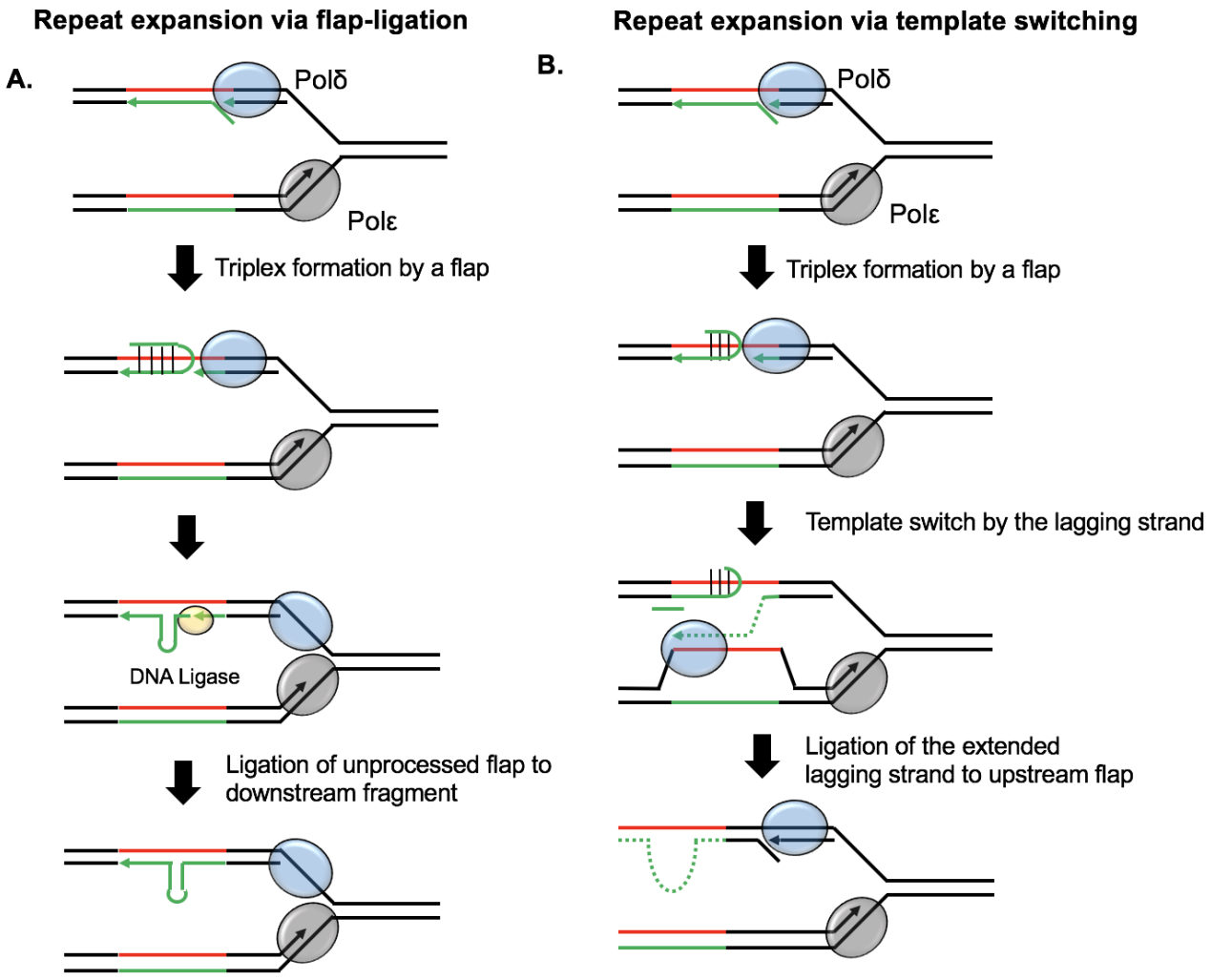


Figure 1. Two models of GAA repeat expansion. **A.** Flap-ligation model of repeat expansion. An unprocessed flap folds back into triplex formation and is then ligated to the 3' end of the downstream flap. In the next round of replication, this unprocessed flap becomes an expansion. **B.** Template switch model of repeat expansion. An unprocessed flap folds back into triplex formation, impeding lagging strand synthesis and triggering the template switch pathway. Pol δ invades the nascent leading strand out-of-register, resulting in an expansion upon the next round of replication.

Importantly, mutations in *RAD27* have been shown to cause a significant increase in GAA repeat expansion (Tsutakawa et al. 2017). The expansion of GAA repeats causes the genetic disease Friedreich's ataxia (FRDA), and thus the mechanisms of GAA repeat expansion are of great scientific importance. The mechanism of these flap-associated expansions is still

unknown; however, we have proposed two possible mechanisms for the process: a flap ligation model and a template switch model.

Flap-ligation model of repeat expansion

In 1997, Tishkoff et al. observed a particular type of mutation that occurred specifically in *rad27* mutant strains: a duplication of sequences flanked by two short repeats of 3-12 bp. They proposed a model for these mutations, wherein an unprocessed flap was ligated to the downstream Okazaki fragment, duplicating the part of the sequence contained in the flap.

Inspired by this study, Gordenin et al. 1997 proposed a flap-ligation model as a mechanism for repeat expansion (Figure 1A). In this model, the 5' end of an unprocessed flap is ligated to the 3' end of the downstream Okazaki fragment. Upon the next round of replication, this flap becomes an expansion. The ligation of the unprocessed flap could be promoted by the tendency of GAA repeats to form an alternate DNA structure, like H-DNA (triplex DNA). FEN1 can cleave only single-stranded flaps and would not be able to process a flap that had folded back into a triplex (Lieber 1997).

Mutations in *rad27* would increase the prevalence of unprocessed flaps, thereby increasing the likelihood of repeat expansion by flap-ligation. In this model, the length of a repeat expansion would depend on the length of an Okazaki fragment flap. The median length of a flap in *S. pombe* was found to be 41 nt in wild-type strains and 89 nt in *fen1Δ* strains (Liu et al. 2017). Thus, we hypothesized that the median expansion length in *rad27* mutants would be ~30 repeats, if the flap ligation model were correct and if flap length is consistent between *S. pombe* and *S. cerevisiae*.

Template switch model of repeat expansion

Post replicative repair (PRR) is a collection of pathways that describe how the cell responds when lagging and leading strand synthesis uncouples. If DNA polymerase encounters a lesion on the DNA that it cannot bypass, it dissociates from PCNA, the sliding clamp that attaches the polymerase to the DNA. Rad6 and Rad18 then ligate a ubiquitin molecule to the

K164 residue of PCNA, which initiates PRR (Hoege et al. 2002). From here, the cell can follow two different pathways: translesion synthesis or template switch.

Translesion synthesis utilizes low-fidelity polymerases, like Pol ζ and Polη, that are capable of synthesizing over lesions (Gao et al. 2017). This pathway has been shown to bypass lesions like thymine dimers, methylated bases, interstrand crosslinks, and DNA-protein crosslinks (Boiteux & Jinks-Robertson 2013). Translesion synthesis is considered an "error-prone" pathway, as the use of low-fidelity polymerases results in an increased frequency of mutations (Gao et al. 2017).

In contrast, the template switch pathway is considered "error-free". This pathway is initiated when a complex of Rad5, Mms2 and Ubc13 poly-ubiquitinates the K164 residue of PCNA. DNA polymerase then switches to using the nascent leading or lagging strand as a template. Once it bypasses the lesion, DNA polymerase returns to the original template. Using an identical strand of DNA as a template theoretically allows this pathway to proceed without error.

Shishkin et al. 2009 first proposed template switching as a mechanism for large-scale GAA repeat expansion, after observing a significant decrease in expansion rate upon knockout of the *RAD5* gene (Figure 1B). In a tract of GAA repeats, an unprocessed flap could fold back to form a triplex structure (H-DNA), which would then act as a lesion that uncouples DNA polymerase from PCNA and initiates the template switch pathway. Given the repetitive nature of GAA repeats, DNA polymerase can then invade the nascent leading strand "out-of-register", resulting in either an expansion or a contraction. In this model, the maximum length of a repeat expansion would be double the original starting length. Large-scale GAA expansions are of particular interest, as in FRDA, the scale of expansions can be quite large. Expanded tracts of GAA repeats of over 1,000 units have been observed, as well as meiotic expansions of up to 208 units (Filla et al. 1996).

WT and *rad27* mutant strains exhibited larger-scale GAA expansion sizes more consistent with the template switch model (Tsutakawa et al. 2017). In addition, whole genome sequencing of *rad27Δ* strains revealed mutations that were identified as template switch events (Omer et al. 2017). 2D gels of plasmids containing GAA repeats revealed X-shaped structures characteristic of template switch intermediates (Follonier et al. 2013, Giannattasio et al. 2014). Lastly, there is chronic poly-ubiquitination of PCNA at the K164 residue in *rad27Δ* strains (Becker et al. 2015). Thus, we hypothesized that flap-associated expansions proceed by the template switch mechanism.

Goals of this project

The goal of this project was to gain further insight into the mechanism of flap-associated GAA repeat expansion. We hypothesized that these expansions proceed via a template switch pathway. To test this hypothesis, we knocked out the *RAD5* gene in *rad27* mutants, thereby eliminating the template switch pathway in these strains. We predicted that knocking out *RAD5* would rescue the expansion-prone phenotype observed in *rad27* mutants and reduce the size of expansions. We also cloned a selectable cassette that would allow us to measure the rate of only large-scale expansions, and we predicted that knocking out *RAD5* would induce an even larger decrease of expansion rate in *rad27* mutants.

We observed that knocking out the *RAD5* gene in *rad27* mutants indeed resulted in a 2-fold reduction in expansion rate for all-scale expansions. However, we did not observe the same effect when measuring only large-scale expansions. Thus, we concluded that Rad5 actually induces small-scale flap-associated expansions. To explain this observation, we have proposed possible template switch and fork regression mechanisms. We also concluded that large-scale flap-associated expansions are Rad5-independent. Consequently, we have proposed a model of large-scale flap-associated expansions based on break-induced replication (BIR).

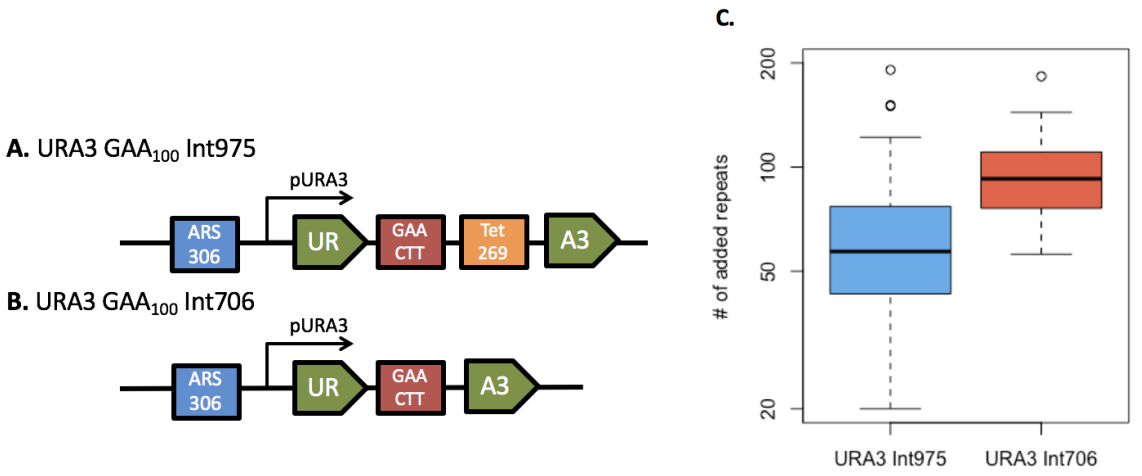


Figure 2. A. URA3 GAA_{100} Int975 cassette containing 100 GAA repeats in a 975 bp long artificial intron. The intron includes a 269 bp segment of a bacterial tetracycline resistance gene. **B.** URA3 GAA_{100} Int706 cassette containing 100 GAA repeats in a 706 bp long artificial intron. **C.** Distribution of expansion lengths observed in both cassettes.

Yeast Experimental Cassette (Figure 2)

To monitor the rate of GAA_{100} expansion in our strains, we used two different selectable cassettes. These cassettes function according to the same principle as those used in Chapter 1: expansion events render cells Ura^- and $5\text{-}FOA^r$ and can thus be selected for by plating on 5-FOA. However, it should be noted that these cassettes contain the endogenous *URA3* promoter, instead of the *GAL1* promoter.

To allow us to measure the rate of only large-scale repeat expansions, we cloned a URA3 Int706 cassette (Figure 2B), which contains a shorter artificial intron than the original URA3 Int975 cassette. In the original creation of the URA3 Int975 cassette, a 269 bp segment from a bacterial tetracycline resistance gene (chosen for a lack of homology to the *S. cerevisiae* genome) was inserted into the intron using an *SphI* site (Shah et al. 2014). Thus, the Tet269 segment was flanked by *SphI* sites. To create the URA3 Int706 cassette, we digested the pYes3-T269-GAA100 plasmid (described in Shah et al. 2014) with *SphI*, gel-purified the solution to remove the Tet269 fragment, then re-ligated the free ends of the cut plasmid.

The removal of 269 bp from the artificial intron increases the threshold length of expansion size required to inactivate the URA3 gene. As seen in Figure 1C, the minimum expansion length detected by the cassette increases from approximately 20 repeats to 60 repeats upon removal of the Tet269 segment. The inactivation of the URA3 gene can be caused by a single-large scale expansion or multiple small-scale expansions.

Mutant strains

For our experiments, we utilized two Rad27 hypomorphs described in Tsutakawa et al. 2017 (Table 1). The mutated amino acid residues impede the DNA threading capability of the enzyme. Genetic data show that these mutations decrease the effectiveness of Rad27, with 4A being less effective than 2E (Tsutakawa et al. 2017). (From this point forward, *rad27* 2E and 4A mutant strains will be referred to simply as 2E or 4A.)

Strain	Mutations
rad27 - 2E	R105E K130E
rad27 - 4A	R104A R105A R127A K130A

Table 1. Point mutations exhibited by the *rad27* mutant strains used in this study. In the 2E mutant, two positively charged residues are converted to two negatively charged residues. In the 4A mutant, four positively charged residues are converted to four neutral residues.

To create the *rad27* mutant strains, we transformed the parent strain CH1585 (*MAT***a**, *leu2-Δ1, trp1-Δ63, ura3-52, his3-200*) with a PCR product containing the desired mutations as well as a downstream *hphMX4* hygromycin resistance marker. As a control, all non-Rad27 mutant strains used in this experiment also contain the downstream hygromycin resistance marker. In Tsutakawa et al. 2017, the presence of this marker was not observed to affect GAA expansion rate.

Fluctuation assay

The protocol for the fluctuation assay for these experiments was identical to that from Chapter 1 with some small changes. Firstly, because the gene cassettes used in these experiments contained a *URA3* promoter instead of a *GAL1* promoter, the strains were plated

directly onto YPD to begin the fluctuation test (omitting the preliminary stage of growth in liquid YPRaffinose).

In addition, *rad27* mutants exhibit an increased rate of GAA length instability. For this reason, we performed a PCR analysis on the colonies after the 40 hour growth stage on YPD to confirm a true starting length of 100 GAA repeats. Colonies that did not exhibit a starting length of 100 GAA repeats were excluded from the data.

For the selective phases, the URA3 Int706 cassette was plated on 0.08% 5-FOA, and the URA3 Int975 cassette was plated on 0.095% 5-FOA.

Results

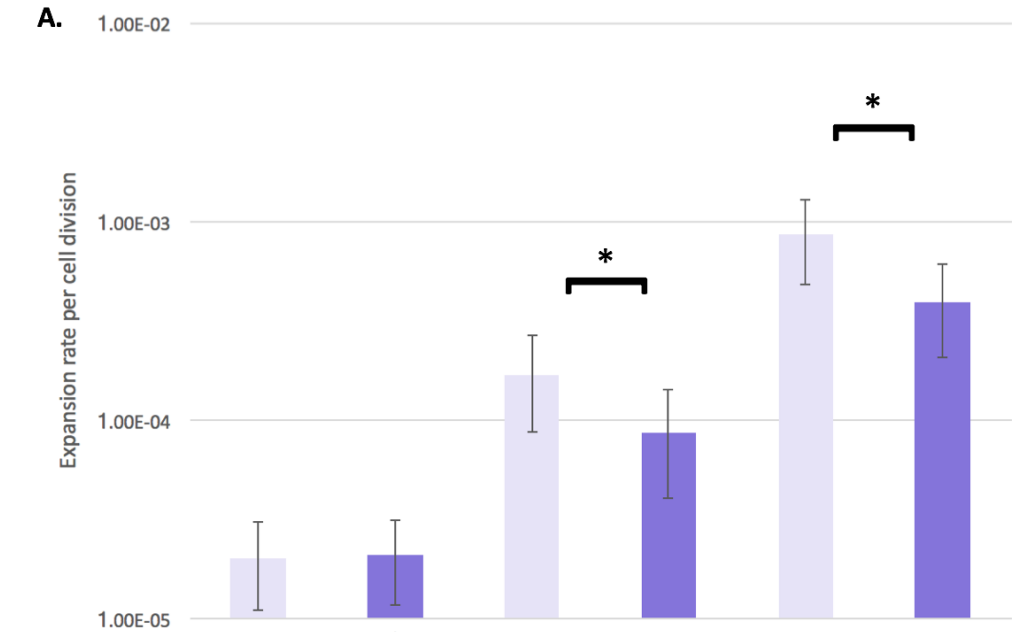
Knocking out RAD5 induces a significant decrease in expansion rate in *rad27* mutants

We first performed fluctuation tests on WT and mutant strains containing the URA3 Int975 cassette to capture a full spectrum of expansion sizes (Figure 3).

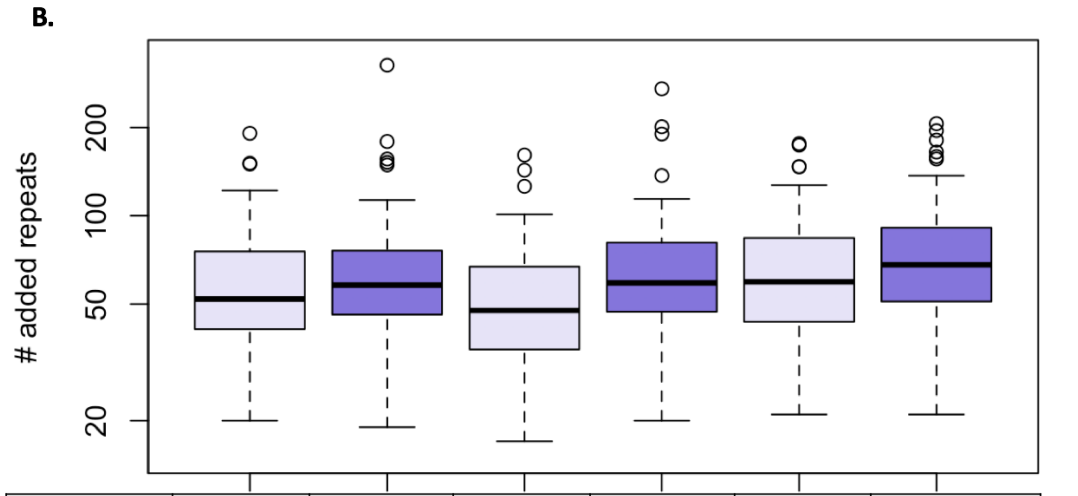
Consistent with the results observed in Tsutakawa et al. 2017, the single 2E and 4A mutants exhibited 10-fold to 50-fold increases in expansion rate over that of the WT. There was no effect of the *RAD5* knockout in the WT strain.

As predicted, the *rad27;rad5Δ* double mutants exhibited significantly decreased expansion rates than the *rad27* single mutants. The effect was consistent across both *rad27* mutants: there was approximately a 2-fold drop in expansion rate upon knocking out *RAD5* in both 2E and 4A strains. This suggests that there are multiple mechanisms of flap-associated repeat expansion, one of which is dependent on Rad5 and is responsible for half of all expansion events. In addition, this Rad5-dependent pathway is not a significant contributor to expansion rate in the WT, as the *RAD5* knockout had no effect on WT.

We also quantified the expansion lengths observed in each strain (Figure 3B). We predicted that we would see shorter expansion sizes in the *rad5Δ* strains. We performed two-sided Kolmogorov-Smirnov tests on the following comparisons and did not observe any significant differences in expansion length: WT vs. *rad5Δ*, 2E vs. 2E;*rad5Δ*, 4A vs. 4A;*rad5Δ,* WT vs. 2E, WT vs. 4A. In fact, the *RAD5* knockout appears to consistently increase the median expansion length in each condition, though not significantly.

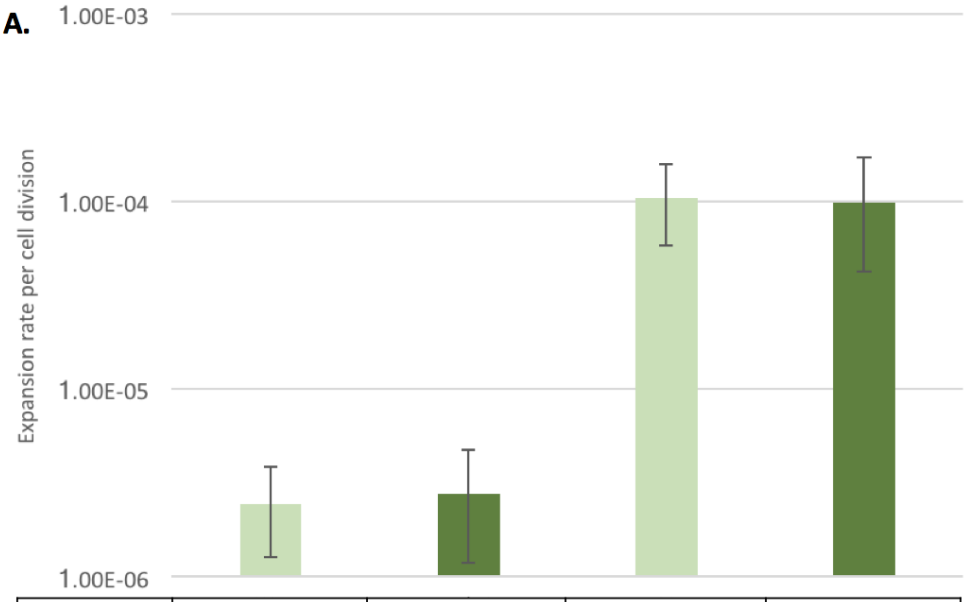


	WT	*rad5Δ*	2E	2E;*rad5Δ*	4A	4A;*rad5Δ*
Exp rate	1.89E-05	2.06E-05	1.69E-04	8.55E-05	8.48E-04	3.89E-04

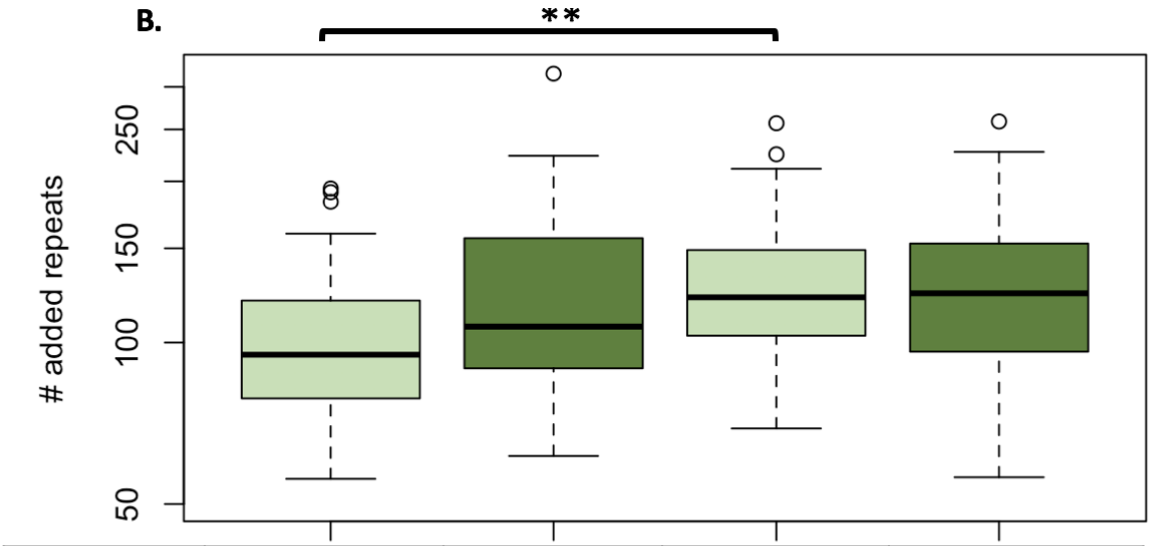


	WT	*rad5Δ*	2E	2E;*rad5Δ*	4A	4A;*rad5Δ*
median # added repeats	52	58	48	59	59	68

Figure 3. Fluctuation test performed on WT and mutant strains containing URA3 Int975 cassette **A.** Rate of GAA repeat expansion per cell division. Error bars indicate 95% confidence intervals. Exact expansion rates are shown below the graph. * indicates $p < 0.02$ as determined by the likelihood ratio test. **B.** Distribution of expansion lengths observed in experimental strains. Median expansion lengths are shown below the graph.



	WT	*rad5Δ*	4A	4A;*rad5Δ*
Exp rate	2.43E-06	2.76E-06	1.04E-04	9.87E-05



	WT	*rad5Δ*	4A	4A;*rad5Δ*
median # added repeats	95	107	122	124

Figure 4. Fluctuation test performed on WT and mutant strains containing URA3 Int706 cassette. **A.** Rate of GAA repeat expansion per cell division. Error bars indicate 95% confidence intervals. Exact expansion rates are displayed below the graph. **B.** Distribution of expansion lengths observed in experimental strains. Median expansion lengths are shown below the graph. ** indicates $p < 0.002$ as determined by a two-sided Kolmogorov-Smirnov test.

Large-scale flap-associated expansions are independent of Rad5

We performed fluctuation tests on strains containing the URA3 Int706 cassette, which detects only expansions longer than ~60 repeats (Figure 4). This expansion event can be caused by a single large-scale expansion event or multiple small- or large-scale events. We hypothesized that we could observe a stronger effect of the *RAD5* knockout on *rad27* mutants when measuring only expansions >60 repeats.

Again, there was no effect of *RAD5* knockout on the WT strain (Figure 4A). In addition, the 4A mutant exhibited a significant increase in expansion rate compared to the WT, as it also did in the URA3 Int706 cassette (Figure 3A).

Since we hypothesized that the large-scale flap-associated expansions were caused by template switching, we predicted that we would observe an even more significant decrease in expansion rate in the 4A;*rad5Δ* strain compared to the 4A strain. Surprisingly, we observed that the partial rescue effect of the *RAD5* knockout on the 4A mutant disappears in the URA3 Int706 cassette (Figure 4A). Again, this result suggests that there are multiple mechanisms of flap-associated repeat expansion, likely one for large-scale expansions and another for small-scale expansions. Furthermore, we can conclude that large-scale flap-associated repeat expansions are not dependent on Rad5.

We also quantified the expansion sizes from this experiment (Figure 4B). In the WT, the median expansion size was 95 repeats, a 43 repeat increase from the median size in the URA3 Int975 cassette. This is consistent with our observation that the minimum expansion length detected by the Int706 cassette is ~40 repeats longer than the minimum length detected by the Int975 cassette (Figure 1C).

We performed two-sided Kolmogorov-Smirnov tests on the following comparisons for the Int706 cassette: WT vs. rad5, WT vs. 4A, *rad5Δ* vs 4A;*rad5Δ*, 4A vs. 4A;*rad5Δ*. Again, we did not observe any significant changes in expansion length in either WT or 4A strains upon knockout of *RAD5*. This is consistent with our conclusion that large-scale expansions are Rad5-

independent. However, we did observe that the 4A mutant had significantly longer expansion sizes than that of the WT. We did not observe this significant difference in strains containing the URA3 Int 975 cassette (Figure 3B).

This result could be a consequence of the increase in expansion rate caused by the 4A mutation. In expansions >100 repeats, the original repeat length has been more than doubled. Thus, these expansions are likely the result of successive expansion events i.e. a cell experiences an expansion event and then one of its descendents experiences another expansion event. Furthermore, these expansions >100 repeats are likely caused by multiple large-scale expansion events rather than multiple short-scale events. The odds of two large-scale expansion events occurring successively are exponentially higher than the odds of four or five small-scale expansion events occuring successively.

Given that expansion rates are often very low, the odds of successive expansion events creating very long expansions is typically also very low. However, because the 4A mutation increases expansion rate over the WT by such a large factor (~45-fold in all-scale and large-scale expansions), it also exponentially increases the odds of successive expansion events, resulting in the observed increase in expansion sizes. This difference between 4A and WT was presumably masked in the URA3 Int975 cassette by the inclusion of small-scale expansions (Figure 3B). In addition, the fact that the 4A and 4A*;rad5Δ* mutants exhibit such large expansion sizes suggests that the mechanism for large-scale expansions rapidly accelerates with longer repeat tracts.

Discussion

Template switch model for small-scale repeat expansions

While we did not anticipate that template switch would be a mechanism for small-scale expansions, this result is not inexplicable. In our original template switch model, an expansion is caused when the lagging strand polymerase switches to the nascent leading strand out-of-register (Figure 1B). It could simply be that this out-of-register invasion is of a smaller-scale than we anticipated.

It is still uncertain how exactly the lagging strand polymerase locates the homologous sequence in the sister chromatid to begin synthesis during template switch. One study by Giannattasio et al. 2014 in *S. cerevisiae* used electron microscopy to analyze template switch intermediates and suggested the following mechanism for this process: The ssDNA gap on the strand containing the lesion binds to the homologous sequence on the DNA duplex in the sister chromatid. The nascent strand then dissociates from this structure, so that it can act as a template for the other nascent strand. According to this model, for an expansion to occur, the DNA duplex would bind to the ssDNA gap out-of-register so that DNA polymerase begins synthesis out-of-register.

The median length of the ssDNA gap on the lesion-containing strand was observed to be 196 nt, and the median length of the three-stranded complex between the two sister chromatids was observed to be 83 nt (Giannattasio et al. 2014). Given these lengths, the DNA duplex could bind out-of-register to the ssDNA gap by a maximum of 113 nt, or 37 repeats. Thus, these EM data predict that the median expansion size caused by template switching would be around 30-40 repeats, which is consistent with our conclusion that small-scale expansions are dependent on Rad5.

Replication fork regression model for short-scale flap-associated expansions

On the other hand, our data could also point to a replication fork regression model of repeat expansion (Figure 5A). In addition to its role as a ubiquitin ligase, Rad5 also has a helicase domain that has been shown to reverse replication forks *in vitro* (Blastyák et al. 2007).

In the process of replication fork regression, the nascent strands are unwound, the parent strands are rewound, and the two nascent strands are then rewound together, forming a four-stranded “chicken-foot” structure (Neelsen & Lopes 2015).

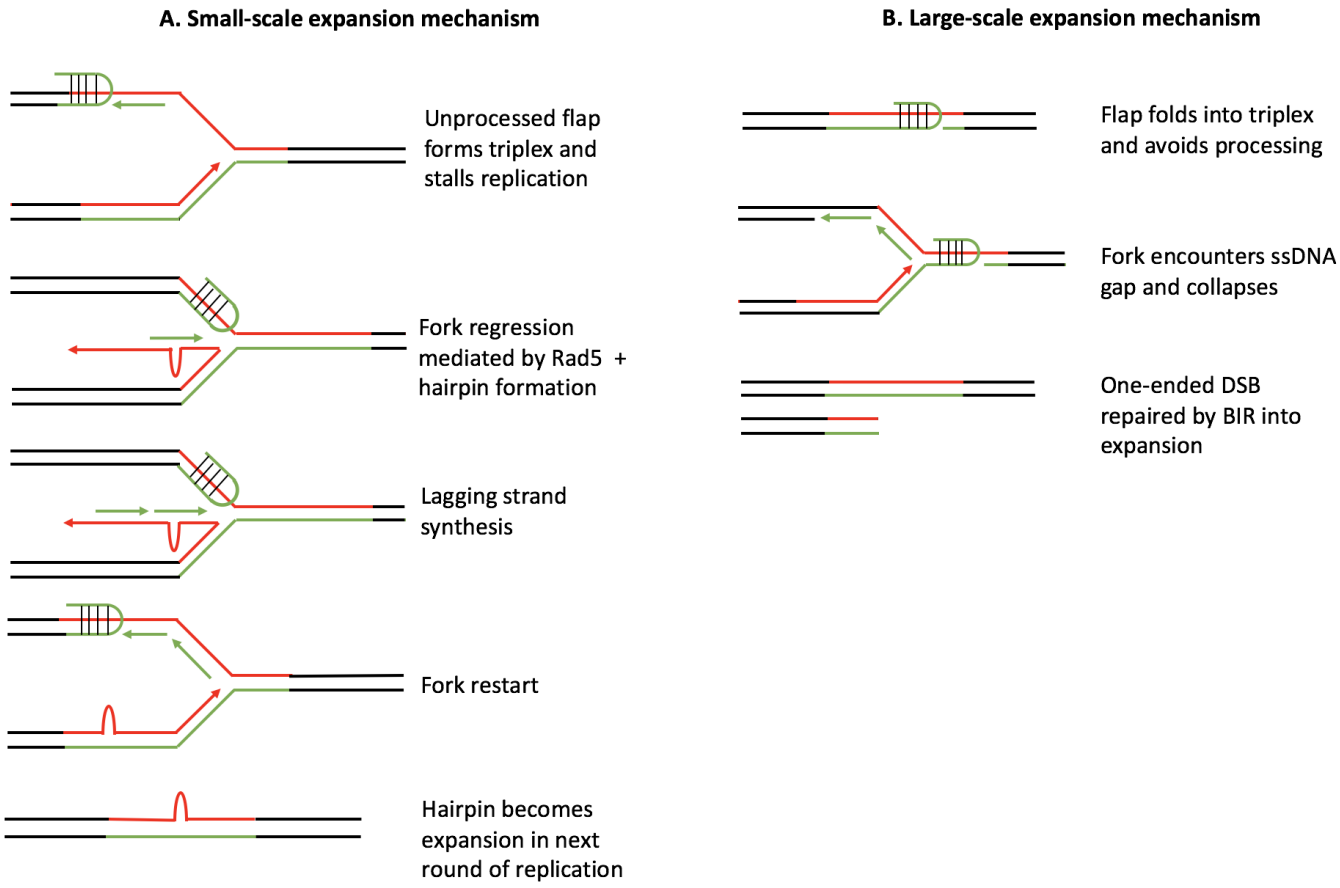


Figure 5. Possible mechanisms for flap-associated GAA repeat expansions. **A.** Short-scale expansion mechanism based on Rad5’s fork regression activity. Triplex formation by an unprocessed Okazaki flap stalls replication and triggers PCNA K164 poly-ubiquitination. Rad5 induces fork regression, creating a hairpin on the nascent leading strand. Lagging strand synthesis may proceed using the nascent leading strand as a template. The hairpin persists through fork restart and becomes a small-scale expansion upon the next round of replication. **B.** Large-scale expansion mechanism based on BIR. An Okazaki flap containing TTC repeats folds into a triplex structure and avoids processing, leaving an adjacent ssDNA gap. Upon the next round of replication, the ssDNA gap collapses the replication fork into a one-ended DSB that is repaired by BIR into a large-scale expansion.

It is still unclear exactly what role replication fork regression plays in the template switch pathway. In some models, fork regression is the mechanism by which the template switch occurs; when the nascent leading and lagging strands are annealed to each other, synthesis can then continue using either of the strands as a template (Blastyák et al. 2007, Meng & Zhao

2016). It is unclear whether fork regression occurs in all template switch events or only in some (Ball et al. 2014).

Replication fork regression in conjunction with hairpin formation has previously been proposed as a model for trinucleotide repeat expansion (Mirkin 2007, Follonier et al. 2013). Furthermore, GAA repeats were shown to induce replication fork regression in SV40 plasmids transfected into human cells (Follonier et al. 2013).

Our fork regression model begins in the same way as our template switch model: triplex formation by a 5' flap containing TTC repeats stalls replication, which induces poly-ubiquitination of PCNA K164. Rad5 induces fork regression but misaligns the nascent leading and lagging strands and creates a hairpin on the nascent leading strand. Lagging strand synthesis may then continue using the nascent leading strand as a template. The hairpin will persist through fork restart and becomes an expansion upon the next round of replication.

This model is consistent with our observation that Rad5-dependent expansions are small-scale. GAA repeat tracts of 99 bp (33 repeats) have been shown to form stable hairpins (Heidenfelder et al. 2003). Thus, expansions caused by fork regression would be small-scale. This model is also consistent with our observation that the *RAD5* knockout had no effect in WT. Fork reversal is rarely observed in WT yeast strains, indicating that it is a mostly pathological phenomenon and only arises during severe replication stress (Neelsen & Lopes 2015).

To distinguish between the template switch and the replication fork regression models, we could utilize separation of function mutants of Rad5, wherein either the ubiquitin ligase domain (r*ad5-I916A*) or the helicase domain (*rad5-QD*) is inactivated (Ball et al. 2014, Choi et al. 2015). If the small-scale Rad5-dependent expansions are caused by template switch, then we would expect the *rad5-I916A* mutant to exhibit the same rescue of expansion rate in the 4A mutant with the URA3 Int975 cassette and no effect from *rad5-QD* (Figure 3A). On the other hand, if the Rad5-dependent expansions are caused by fork regression, then we would expect the opposite.

Break-induced replication as a possible mechanism for large-scale flap-associated expansions

Now that we have shown large-scale flap-associated repeat expansions to be Rad5-independent, we must posit a non-template switch mechanism for large-scale expansions. Some likely candidates could be other pathways of recombinational repair.

Consequently, we propose a model for large-scale flap-associated expansions based on break-induced replication (BIR) (Figure 5B). BIR is initiated after a one-ended double strand break (DSB). During BIR, the 3' end of a one-ended DSB will invade the homologous chromosome to use as a template to repair the break. In repetitive regions, this strand invasion could theoretically occur "out-of-register" and result in an expansion. An out-of-register invasion during BIR has been proposed as a mechanism for large-scale CAG and GAA repeat expansions (Kim et al. 2016, Neil et al. 2018).

A *rad27* deficiency could result in an increased frequency of one-ended DSBs. If the flap of an Okazaki fragment were to escape processing, there would remain a section of ssDNA adjacent to the flap. Consequently, upon the next round of replication, this ssDNA gap could collapse the replication fork and cause a one-ended DSB.

Liu et al. 2017 observed that the increase in RPA foci observed in a *fen1Δ* strain compared to the WT persisted until the G1 phase of mitosis, likely due to the binding of RPA to unprocessed flaps. This result suggests that unprocessed flaps can persist until the next round of replication. Moreover, an unprocessed flap containing TTC repeats (as is in our system) would be particularly difficult to process if it formed into a triplex. If a 5' flap grows longer than ~30 nt due to a failure of Rad27 to cleave, it will enter the long flap maturation pathway (Burgers 2008). In this pathway, the RPA binds to the ssDNA flap to signal for Dna2, another exonuclease, to cleave the flap (Zaher et al. 2018). Replication Protein A (RPA) is a protein that binds to ssDNA during DNA replication and repair to prevent genome instability (Zou et al.

2006). However, a triplex forming from an Okazaki flap would theoretically exclude most RPA binding due its lack of ssDNA and consequently resist removal.

Furthermore, *rad27Δ* is inviable when combined with *rad51Δ, rad52Δ*, or *pol32Δ* mutations (Tong et al. 2001, Tong et al. 2004). We also found 4A;*rad51Δ* and 4A;*rad52Δ* double mutants to be inviable (data not shown). Pol32 is a subunit of Pol δ that is required for most cases of BIR, and Rad51 and Rad52 are enzymes that participate in strand invasion and exchange in various pathways of recombination (Anand et al. 2013). Thus, in the absence of recombinational repair mechanisms, *rad27* mutant strains cannot survive. The reason could be that *rad27* mutants accumulate ssDNA gaps that collapse replication forks into DSBs and consequently require recombination to repair.

However, it is also possible that these synthetic lethalities have other causes. For example, the synthetic lethality of *rad27Δ;pol32Δ* could be due to Pol32's role in Okazaki fragment synthesis instead of its role in BIR. In the absence of Rad27, flaps are cleaved by Dna2, but Dna2 can only cleave long flaps (Zaher et al. 2018). Furthermore, Pol32 has strand displacement activity and consequently promotes long flap formation during Okazaki fragment synthesis (Burgers 2008). Therefore, in the absence of Pol32, long flaps cannot form, and in the absence of Rad27, short flaps cannot be cleaved, resulting in the synthetic lethality of *rad27Δ;pol32Δ* strain. In addition, the synthetic lethality of *rad27Δ;rad51Δ* and *rad27Δ;rad52Δ* could be due to the roles of Rad51 and Rad52 in PRR. Impaired flap processing in *rad27* mutants could result in an accumulation of ssDNA gaps (Callahan et al. 2003, Becker et al. 2015). These gaps would then be repaired a recombinational repair pathway in PRR that uses Rad51 or Rad52 (Xu et al. 2015). Thus, in *rad27Δ;rad51Δ and rad27Δ;rad52Δ* strains, the cell accumulates ssDNA gaps, which cannot be repaired via PRR, resulting in synthetic lethality.

It should also be noted that this proposed BIR-based mechanism would occur only in *rad27* mutants and not WT strains, as single mutants of these BIR enzymes do not exhibit rescues in GAA expansion rate compared to the WT: *rad51Δ* and *pol32Δ* single mutants

exhibited the same expansion rates as WT strains, and *rad52Δ* strains exhibited slightly increased expansion rates compared to WT strains (Neil et al. 2018, McGinty et al. 2017).

With this BIR-based mechanism, expansion rate would accelerate with longer repeats, as was indicated by the expansion sizes of the URA3 Int706 cassette (Figure 4B). The longer the repeat, the more Okazaki 5' flaps will contain GAA repeats, form triplexes, and escape processing. Thus, replication fork collapse and BIR becomes much more likely with longer repeats.

We can test this model by knocking out enzymes involved in BIR. Though *rad27Δ*;*pol32Δ* strains are inviable, it still may be possible for 4A;*pol32Δ* mutants to be viable. We could also knock out Pif1, a helicase that participates in BIR (Buzovetsky et al. 2017). According to our BIR-based model for large-scale expansions, we would expect 4A;*pif1Δ* and 4A;*pol32Δ* mutants to exhibit strong rescues of expansion rate compared to 4A single mutants when looking only at large-scale expansions (URA3 Int706 cassette) and partial rescues when looking at all-scale expansions (URA3 Int975 cassette).

In conclusion, we have found that small-scale flap-associated GAA repeat expansions are Rad5-dependent. We propose that these small-scale expansions are induced by out-of-register strand invasion during template switching or by hairpin formation during replication fork regression. We have also found that large-scale flap-associated GAA repeat expansions are Rad5-independent. We propose that these large-scale expansions are induced by unprocessed flaps creating a one-ended DSB upon the next round of replication that is repaired by BIR into an expansion.

Citations

Ball, L. G., Xu, X., Blackwell, S., Hanna, M. D., Lambrecht, A. D., & Xiao, W. (2014). The Rad5 helicase activity is dispensable for error-free DNA post-replication repair. *DNA Repair, 16*, 74-83. doi:10.1016/j.dnarep.2014.02.016

Becker, J. R., Pons, C., Nguyen, H. D., Costanzo, M., Boone, C., Myers, C. L., & Bielinsky, A. (2015). Genetic Interactions Implicating Postreplicative Repair in Okazaki Fragment Processing. *PLOS Genetics, 11*(11). doi:10.1371/journal.pgen.1005659

Blastyák, A., Pintér, L., Unk, I., Prakash, L., Prakash, S., & Haracska, L. (2007). Yeast Rad5 Protein Required for Postreplication Repair Has a DNA Helicase Activity Specific for Replication Fork Regression. *Mol. Cell, 28*(1), 167-175. doi:10.1016/j.molcel.2007.07.030

Boiteux, S., & Jinks-Robertson, S. (2013). DNA Repair Mechanisms and the Bypass of DNA Damage in Saccharomyces cerevisiae. *Genetics, 193*(4), 1025-1064. doi:10.1534/genetics.112.145219

Burgers, P. M. (2008). Polymerase Dynamics at the Eukaryotic DNA Replication Fork. *J. Biol. Chem., 284*(7), 4041-4045. doi:10.1074/jbc.r800062200

Buzovetsky, O., Kwon, Y., Pham, N., Kim, C., Ira, G., Sung, P., & Xiong, Y. (2017). Role of the Pif1-PCNA Complex in Pol delta dependent Strand Displacement DNA Synthesis and Break Induced Replication. *Cell Rep., 21*(7), 1707-1714. doi:10.2210/pdb6b8i/pdb

Callahan, J. L., Andrews, K. J., Zakian, V. A., & Freudenreich, C. H. (2003). Mutations in Yeast Replication Proteins That Increase CAG/CTG Expansions Also Increase Repeat Fragility. *Mol. Cell. Biol., 23*(21), 7849-7860. doi:10.1128/mcb.23.21.7849-7860.2003

Chapados, B. R., Hosfield, D. J., Han, S., Qiu, J., Yelent, B., Shen, B., & Tainer, J. A. (2004). Structural Basis for FEN-1 Substrate Specificity and PCNA-Mediated Activation in DNA Replication and Repair. *Cell, 116*(1), 39-50. doi:10.1016/s0092-8674(03)01036-5

Choi, K., Batke, S., Szakal, B., Lowther, J., Hao, F., Sarangi, P., . . . Zhao, X. (2015). Concerted and differential actions of two enzymatic domains underlie Rad5 contributions to DNA damage tolerance. *Nucleic Acids Res., 43*(5), 2666-2677. doi:10.1093/nar/gkv004

Filla, A., Michele, G. D., Cavalcanti, F., Pianese, L., Monticelli, A., Campanella, G., & Cocozza, S. (1996). The relationship between trinucleotide (GAA) repeat length and clinical features in Friedreich ataxia. *Am. J. Hum. Genet.,* 554-560.

Follonier, C., Oehler, J., Herrador, R., & Lopes, M. (2013). Friedreich's ataxia–associated GAA repeats induce replication-fork reversal and unusual molecular junctions. *Nat. Struct. Mol. Biol., 20*(4), 486-494. doi:10.1038/nsmb.2520

Gary, R., Kim, K., Cornelius, H. L., Park, M. S., & Matsumoto, Y. (1999). Proliferating Cell Nuclear Antigen Facilitates Excision in Long-patch Base Excision Repair. *J. Biol. Chem., 274*(7), 4354-4363. doi:10.1074/jbc.274.7.4354

Gao, Y., Mutter-Rottmayer, E., Zlatanou, A., Vaziri, C., & Yang, Y. (2017). Mechanisms of Post-Replication DNA Repair. *Genes, 8*(12), 64. doi:10.3390/genes8020064

Giannattasio, M., Zwicky, K., Follonier, C., Foiani, M., Lopes, M., & Branzei, D. (2014). Visualization of recombination-mediated damage bypass by template switching. *Nat. Struct. Mol. Biol., 21*(10), 884-892. doi:10.1038/nsmb.2888

Gordenin, D. A., Kunkel, T. A., & Resnick, M. A. (1997). Repeat expansion — all in a flap? *Nat. Genet., 16*(2), 116-118. doi:10.1038/ng0697-116

Hoege, C., Pfander, B., Moldovan, G., Pyrowolakis, G., & Jentsch, S. (2002). RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. *Nature, 419*(6903), 135-141. doi:10.1038/nature00991

Kim, J. C., Harris, S. T., Dinter, T., Shah, K. A., & Mirkin, S. M. (2016). The role of break-induced replication in large-scale expansions of $(CAG)_n/(CTG)_n$ repeats. *Nat. Struct. Mol. Biol., 24*(1), 55-60. doi:10.1038/nsmb.3334

Lieber, M. R. (1997). The FEN-1 family of structure-specific nucleases in eukaryotic DNA replication, recombination and repair. *BioEssays, 19*(3), 233-240. doi:10.1002/bies.950190309

Liu, B., Hu, J., Wang, J., & Kong, D. (2017). Direct Visualization of RNA-DNA Primer Removal from Okazaki Fragments Provides Support for Flap Cleavage and Exonucleolytic Pathways in Eukaryotic Cells. *J. Biol. Chem., 292*(12), 4777-4788. doi:10.1074/jbc.m116.758599

McGinty, R. J., Puleo, F., Aksenova, A. Y., Hisey, J. A., Shishkin, A. A., Pearson, E. L., . . . Mirkin, S. M. (2017). A Defective mRNA Cleavage and Polyadenylation Complex Facilitates Expansions of Transcribed (GAA) n Repeats Associated with Friedreich's Ataxia. Cell Rep., 20(10), 2490-2500. doi:10.1016/j.celrep.2017.08.051

Meng, X., & Zhao, X. (2016). Replication fork regression and its regulation. *FEMS Yeast Res.*doi:10.1093/femsyr/fow110

Mirkin, S. M. (2007). Expandable DNA repeats and human disease. *Nature, 447*(7147), 932-940. doi:10.1038/nature05977

Neelsen, K. J., & Lopes, M. (2015). Replication fork reversal in eukaryotes: From dead end to dynamic response. *Nat. Rev. Mol. Cell. Biol.,16*(4), 207-220. doi:10.1038/nrm3935

Neil, A. J., Liang, M. U., Khristich, A. N., Shah, K. A., & Mirkin, S. M. (2018). RNA–DNA hybrids promote the expansion of Friedreich's ataxia $(GAA)_n$ repeats via break-induced replication. *Nucleic Acids Res*., doi:10.1093/nar/gky099

Omer, S., Lavi, B., Mieczkowski, P. A., Covo, S., & Hazkani-Covo, E. (2017). Whole-Genome Sequence Analysis of Mutations Accumulated in rad27Δ Yeast Strains with Defects in the Processing of Okazaki Fragments Indicates Template-Switching Events. *G3 (Bethesda)*. doi:10.1534/g3.117.300262

Parenteau, J., & Wellinger, R. J. (2002). Differential Processing of Leading- and Lagging-Strand Ends at Saccharomyces cerevisiae Telomeres Revealed by the Absence of Rad27p Nuclease. *Genetics, 162*(4), 1583-1594.

Saharia, A., Guittat, L., Crocker, S., Lim, A., Steffen, M., Kulkarni, S., & Stewart, S. A. (2008). Flap Endonuclease 1 Contributes to Telomere Stability. *Curr. Biol., 18*(7), 496-500. doi:10.1016/j.cub.2008.02.071

Shah, K., Shishkin, A., Voineagu, I., Pavlov, Y., Shcherbakova, P., & Mirkin, S. (2012). Role of DNA Polymerases in Repeat-Mediated Genome Instability. *Cell Rep., 2*(5), 1088-1095. doi:10.1016/j.celrep.2012.10.006

Shah, K. A., McGinty, R. J., Egorova, V. I., & Mirkin, S. M. (2014). Coupling transcriptional state to large-scale repeat expansions in yeast. *Cell Rep.*, *9*(5), 1594–1602. http://doi.org/10.1016/j.celrep.2014.10.048

Shishkin, A. A., Voineagu, I., Matera, R., Cherng, N., Chernet, B. T., Krasilnikova, M. M., Narayanan, V., Lobachev, K., Mirkin, S. M. (2009). Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. *Mol. Cell*, *35*(1), 82–92. http://doi.org/10.1016/j.molcel.2009.06.017

Smith, D. J., & Whitehouse, I. (2012). Intrinsic coupling of lagging-strand synthesis to chromatin assembly. *Nature, 483*(7390), 434-438. doi:10.1038/nature10895

Tishkoff, D. X., Filosi, N., Gaida, G. M., & Kolodner, R. D. (1997). A Novel Mutation Avoidance Mechanism Dependent on S. cerevisiae RAD27 Is Distinct from DNA Mismatch Repair. *Cell, 88*(2), 253-263. doi:10.1016/s0092-8674(00)81846-2

Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Pag, N., … Boone, C. (2001). Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. *Science,294*(5550), 2364-2368. doi:10.1126/science.1065810

Tong, A. H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X. … Boone, C. (2004). Global Mapping of the Yeast Genetic Interaction Network. *Science, 303*(5659), 808-813. doi:10.1126/science.1091317

Tsutakawa, S. E., Thompson, M. J., Arvai, A. S., Neil, A. J., Shaw, S. J., Algasaier, S. I., … Tainer, J. A. (2017). Phosphate steering by Flap Endonuclease 1 promotes 5′-flap specificity and incision to prevent genome instability. *Nat. Comm., 8*, 15855. doi:10.1038/ncomms15855

Xu, X., Blackwell, S., Lin, A., Li, F., Qin, Z., & Xiao, W. (2015). Error-free DNA-damage tolerance in *Saccharomyces cerevisiae*. *Mutat. Res. Rev. Mutat. Res.,764*, 43-50. doi:10.1016/j.mrrev.2015.02.001

Zaher, M. S., Rashid, F., Song, B., Joudeh, L. I., Sobhy, M. A., Tehseen, M., … Hamdan, S. M. (2018). Missed cleavage opportunities by FEN1 lead to Okazaki fragment maturation via the long-flap pathway. *Nucleic Acids Res., 46*(6), 2956-2974. doi:10.1093/nar/gky082

Zheng, L., & Shen, B. (2011). Okazaki fragment maturation: nucleases take centre stage. *J. Mol. Cell. Biol., 3*(1), 23-30. doi:10.1093/jmcb/mjq048

Zheng, L., Jia, J., Finger, L. D., Guo, Z., Zer, C., & Shen, B. (2010). Functional regulation of FEN1 nuclease and its link to cancer. *Nucleic Acids Res., 39*(3), 781-794. doi:10.1093/nar/gkq884

Zou, Y., Liu, Y., Wu, X., & Shell, S. M. (2006). Functions of human replication protein A (RPA): From DNA replication to DNA damage and stress responses. *J. Cell. Physiol., 208*(2), 267-273. doi:10.1002/jcp.20622