The Interaction of DNA Double-Strand Break Repair and CAG/CTG Trinucleotide Repeat Instability in *Drosophila melanogaster*

Jane Blackmer

Senior Honors Thesis

Biology Department, Tufts University

Under the advisement of Dr. Mitch McVey

2018

Table of Contents

Abstract……3

Introduction……4

Trinucleotide Repeats and Their Role in Disease……4

The Molecular Mechanisms of CAG/CTG Trinucleotide Repeat Instability……..8

DNA Double-strand Breaks and Their Role in Repeat Expansion……..15

Goals of This Study……20

Materials and Methods……..21

Fly Crosses……21

Molecular Cloning……22

Scoring of the Assay……27

Synthesis Tract Length PCR……29

Results……..33

Creating the Assay……33

Interpretation and Characterization of the Assay……54

Repair Outcomes of the Assay……59

Discussion……60

An Inter-Homolog Repair Assay Designed for SDSA using TNRs as a Template..60

Molecular Cloning Resulted in Creation of All but Two Plasmids……..61

The Vast Majority of Repair is Comprised of End Joining Events……..62

Future Iterations of the Assay will be Further Biased to Repair Via SDSA……63

Future Directions and Conclusions……..66

Acknowledgements……71

References……72

Supplemental Figures……79

Abstract

Trinucleotide repeats are known to be the underlying molecular cause of over two dozen neurodegenerative disorders in humans. CAG/CTG repeats in particular are the cause of Spinal and Bulbar Muscular Atrophy, multiple types of Spinocerebellar Ataxia, Dentatorubral-Pallidoluysian Atrophy, and, most well-known, Huntington's Disease. A common feature of trinucleotide repeats is their dependence on length. When the repeat number is below a certain threshold, usually less than 40 repeats in humans, depending on the disorder, there are no symptoms of disease. Once expanded past a certain threshold, however, these repeats cause pathology on the DNA, RNA, and protein level. Therefore, the mechanism of expansion has been a popular topic of research over the past decades. Trinucleotide repeats are known to expand in both the somatic tissue and in the germline. Expansions occurring in the germline lead to a phenomenon known as genetic anticipation, where there is a larger number of repeats being passed on to each successive generation, causing an earlier age of onset of the disease as well as a more severe phenotype. There is an observed increase in the instability of CAG trinucleotide repeats in the human male germline, and a paternal bias for the inheritance of expanded repeats. Several models for repeat instability have been proposed. These models focus on the ability of these GC-rich repeats to form secondary structures, especially when the DNA is single stranded. These models therefore describe instability as occurring during replication, transcription, and repair of trinucleotide repeats. Previous studies in yeast have observed an increase in instability of CAG repeats during double-strand break repair via homologous recombination. Therefore, the goal for this experiment was to create and characterize a novel assay in which a double-strand break is created in the male germline of *Drosophila melanogaster* and repaired via homologous recombination (specifically, synthesis-dependent strand annealing) through a region of

CAG/CTG trinucleotide repeats. We created several constructs through molecular cloning in order to study repeat instability with a varying number of repeats, as well as with and without the fluorescent marker, GFP. We injected these constructs into *Drosophila* embryos and created stocks. We performed the first iteration of the assay in a stock containing 71 CAG repeats before the induction of a double-strand break and repair. Results of the assay showed that the majority of repair events were end joining, with a small percentage of repair events being completed synthesis-dependent strand annealing, and an even smaller number of events initiating but later aborting repair via synthesis-dependent strand annealing and completing repair via end joining. Experiments are ongoing to determine both the stability of the repeats across generations, as well as the stability of the repeats after repair via synthesis-dependent strand annealing.

Introduction

Trinucleotide Repeats and Their Role in Disease

Trinucleotide repeats (TNRs) are a type of microsatellite sequence comprised of three repeated nucleotides in the genome. These TNRs are unstable, and their expansion can cause over two dozen diseases, all of which are highly heritable and affect the central nervous system (Budworth and McMurray, 2013; La Spada and Taylor, 2010). There are many microsatellites present in the human genome, and individuals differ in the exact number of repeats they possess. Additionally, not all trinucleotide repeats cause pathology. Many TNRs are actually stable, but the ones that cause disease are unstable and often contract or expand across generations. The disease-causing TNRs are often more GC-rich than their innocuous counterparts (Kuhl et al., 1993).

TNRs vary in their permutation of nucleotides, location in the genome, and the mechanism by which they cause disease. The mechanisms by which TNRs cause disease in

humans can occur on the DNA, RNA, or protein level. The factors that influence instability, and thereby disease, on the DNA level (also called *cis*-acting factors), include the number of trinucleotide repeats, their capacity to form secondary structures (such as cruciforms, G-quadruplex structures, and hairpins), and their orientation in the genome with respect to origins of replication. The *trans*-acting factors include disruptions on the RNA or protein level, such as the propensity for certain repair proteins to cause increased instability of the repeats (reviewed in Lenzmeier et al., 2003). Another *trans*-acting factor of instability are the protein products of the CAG repeats, called polyglutamine (or PolyQ) repeats, as they can affect the protein function of DNA repair machinery (Jung et al., 2011). Additionally, these PolyQ repeats in neurons stand in the way of transport along the axon, leading to many neurological symptoms (Krench et al., 2013).

CAG repeats in particular cause many known neurodegenerative diseases, including Spinal and Bulbar Muscular Atrophy (SBMA), several Spinocerebellar Ataxias (SCA types1-3, 6, 7, 10, 12, and 17), Dentatorubral-Pallidoluysian Atrophy (DRPLA), and Huntington's Disease. CTG repeats also lead to pathology, causing both Huntington disease-like 2 (HDL2) and Myotonic Dystrophy types 1 and 2. Both CAG and CTG repeats cause SCA8 (La Spada and Taylor, 2010). The cause of Huntington's Disease was originally discovered by The Huntington's Disease Collaborative Research Group in 1993, when a polymorphic CAG TNR was found in the exon of one gene on 4p16.3 (huntingtin, or HTT). This group also discovered that the transmission of a more severe disease allele appears to be going through the male germline.

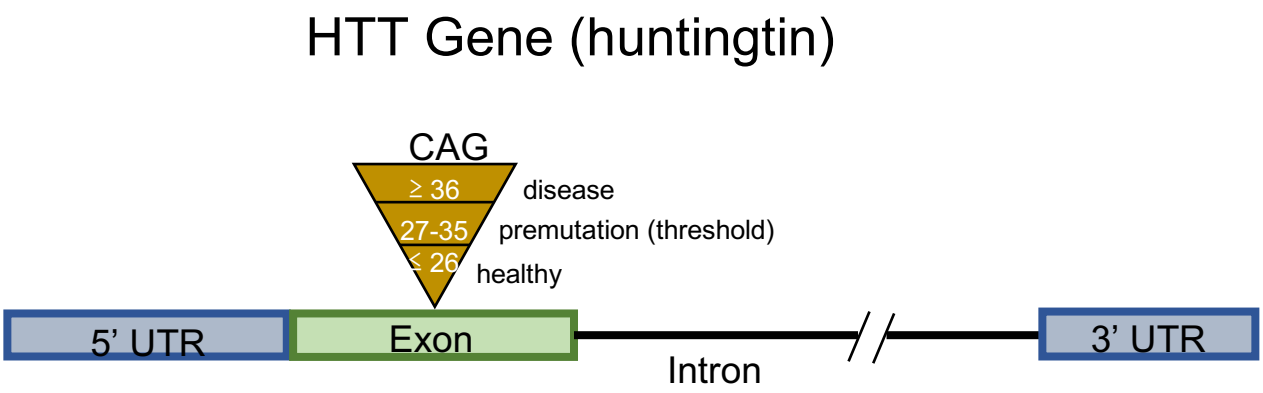


Figure 1. Huntingtin gene (HTT) with varying numbers of repeats (within the green exon) and the expected disease outcomes. In healthy individuals with under 26 repeats, no discernable phenotype is expected. Individuals with 27-35 CAG repeats still show no disease phenotype, but are carriers for the disease, as the repeat number may surpass the disease threshold in the next generation. In individuals with greater than 36 repeats, Huntington's Disease is observed, with the greater the number of repeats producing a more severe phenotype and an earlier age of onset.

Huntington's Disease presents with a number of neurological, psychiatric and motor symptoms, including chorea progressing to bradykinesia, sleep disturbances, and depression, among other symptoms (reviewed in Roos, 2010). *Drosophila* serve as a good model for repeat instability diseases, as CAG/CTG trinucleotide repeats in *Drosophila* manifest with similar symptoms to those in humans. *Drosophila* with either a mutant form of HTT or an RNAi-mediated knockdown of the *htt* gene exhibit sleep disturbances, similar to that observed in human patients (Gonzales et al., 2010). Additionally, the *htt* gene is well conserved across species, and the *Drosophila* HTT protein contains HEAT repeats, which are the regions that allow the protein to interact with other HTT proteins, leading to PolyQ repeat aggregates (Krench et al., 2013).

Similar to in humans, *Drosophila* show a bias for expansions over contractions of repeats that are passed between generations, exhibiting expansions at a three times higher frequency than contractions (Jung et al., 2007). Additionally, *Drosophila* also exhibit increased instability with

an increased repeat length, as has been observed in humans (Jung et al., 2007). In *Drosophila*, a repeat length of CAG78 was fairly stable across generations without transcription. With transcription, there was more instability, including examples of expansions and contractions of greater than 10 repeats within just one generation (Jung et al., 2007). When the repeat number was increased to CAG270, there was greater instability with or without transcription, exhibiting a length dependence on instability. The CAG270 *Drosophila* passed on the repeats maternally, and when there was transcription of the repeats, ~35% of the progeny had a different number of repeats than the mother, ~13% showing +1 expansions, and ~13% showing +2 or more expansions. In the progeny that had come from a stock of flies that were not transcribing the repeat, the overall percentage of +1 expansions was not significantly different but exhibited a smaller number of larger expansions (Jung et al., 2011).

One notable difference between humans and *Drosophila* is the method of inheritance of the repeats. In many mammals, including mice and humans, there is a very strong pattern of inheritance of expansions from the father, and not the mother. The opposite is true in *Drosophila*, where a stronger imprinting effect from the mother is observed. The reasoning for this pattern of male inheritance is thought to be because the male germline cells go through cell division continuously during spermatogenesis, therefore leading to the instability of their TNRs through replication. Conversely, in oogenesis in the female germline, cells are arrested in meiosis I and then again in meiosis II, and go through far fewer divisions than the developing sperm do overall. This leads to fewer expansions being inherited from the mother. Meiotic arrest could give the cell time to contract the repeats and repair any expansions that occurred during replication. Additionally, it is thought that since there are multiple developing oocytes, but only one wins out to be ovulated, perhaps the oocytes with more expansions are degraded and the

oocyte with the fewest repeats survives to become the primary oocyte (Sato et al. 1999). It is hypothesized that because the female germ line cells in *Drosophila* do not experience meiotic arrest, the pattern of paternal transmission is no longer observed, and a maternal bias is present (Jung et al., 2007). However, repeat expansions occurring in the human male germ line have been shown to occur at several different stages, with many of them occurring in the pre-meiotic germline, suggesting that meiosis may not be the main cause of instability. Spermatogonial cells of the human male germline go through many successive divisions over the course of one's lifetime, more so than in *Drosophila* or mice, meaning there is an increased opportunity for expansions to occur pre-meiotically, possibly explaining this distinction (Yoon et al., 2003).

Interestingly, in mice, it has been shown that not only is there a bias towards paternal transmission of TNRs, but there is also a difference in inheritance based off of the sex of the embryo. Female embryos are more prone to contraction and males to expansion. This likely points to a role in the X or Y chromosomes in mediating these differential instability events (Kovtun et al., 2000). This same study also observed that CAG repeats were inherited in a Mendelian ratio, and that while there was a paternal bias of repeat inheritance, the range of repeat number in the offspring was not accurately reflected in the range of repeats in the paternal germ line. These results imply that there are some expansions occurring in the embryos, and once again, there are more expansions occurring in male embryos, implicating a role for the sex chromosomes in repeat instability (Kovtun et al., 2000).

The Molecular Mechanisms of CAG/CTG Trinucleotide Repeat Instability

The larger the number of repeats present in the genome, the more unstable they are, meaning the more prone they are to contract and expand. Once there is a large enough number of repeats, they become premutation alleles, becoming dramatically more unstable and expanding

between generations. While premutation alleles do not cause pathology, they have the ability to expand to the point of disease in the next generation (Kuhl et al., 1993). TNR expansions are thought to be a product of secondary structure formation on either the template or the daughter strand during replication or repair (see figure 2). Different types of trinucleotide repeats can form different secondary structures, with CAG/CTG repeats forming hairpins that are Watson-Crick base paired at the cytosines and guanines, with a mismatch in the middle with either adenine or thymine (McMurray, 1999). While both strands of DNA, either containing the CAG or CTG version of repeats, are able to form hairpins, CTG repeats form more stable hairpins than CAG repeats (Petruska et al., 1996). However, it is more energetically favorable for CAG repeats to slip during replication or repair and form the hairpin in the first place, over CTG repeats (Hartenstine et al., 2000). Replication slippage occurs when the repeats fold over on themselves to form hairpin structures. The longer the repeats, the more hairpins can form, leading to more expansions (Petruska et al., 1996). These hairpins can stall the DNA polymerase, causing it to dissociate from the template strand. When the polymerase dissociates, some of the nascent strand can dissociate as well, leading to reannealing and re-replication of the repeats on the template strand, thus causing an expansion event (Viguera et al., 2001).

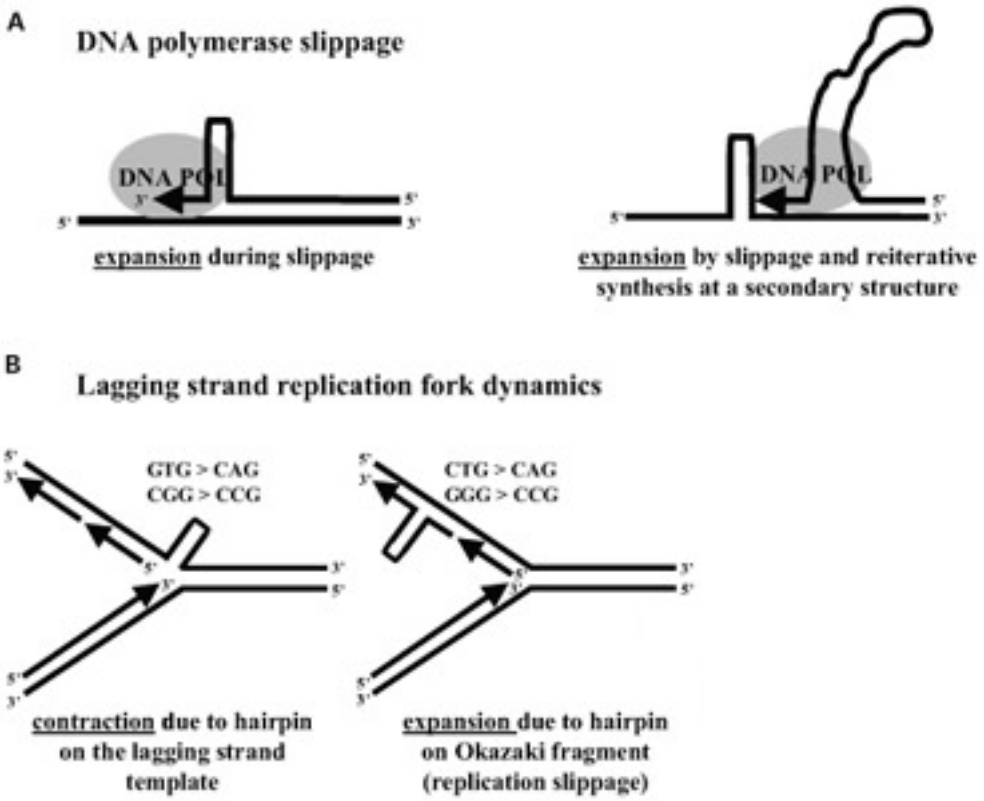


Figure 2. Different types of replication slippage. A) Left: Replication slippage of the nascent strand during replication, leading to an expansion. Right: A hairpin in the template causes polymerase slippage and reannealing to an earlier locus on the template strand, causing the polymerase to re-synthesize through a region of repeats, leading to expansion. B) Left: Hairpin formation on the lagging strand template, leading to contractions. This is more common in CTG repeats than CAG repeats. Middle: Slippage of the lagging daughter strand in replication leads to expansions. Once again, this is more common in CTG repeats than CAG repeats (Figure adapted from Lenzmeier et al., 2003).

Longer tracts of repeats are more prone to instability because they can create a larger number of hairpins (not because they form fewer, larger hairpins, as this would be energetically unfavorable). In yeast, it has been observed that the orientation of repeats in the genome with respect to the direction of replication plays an important role in TNR instability. Because CTG repeats form more stable hairpins, when present on the lagging strand during replication, more instability occurs, with frequent deletions occurring. Conversely, when CAG repeats are present

on the lagging strand, there are fewer changes to repeat number overall, but more expansions than contractions (Freudenreich et al., 1997).

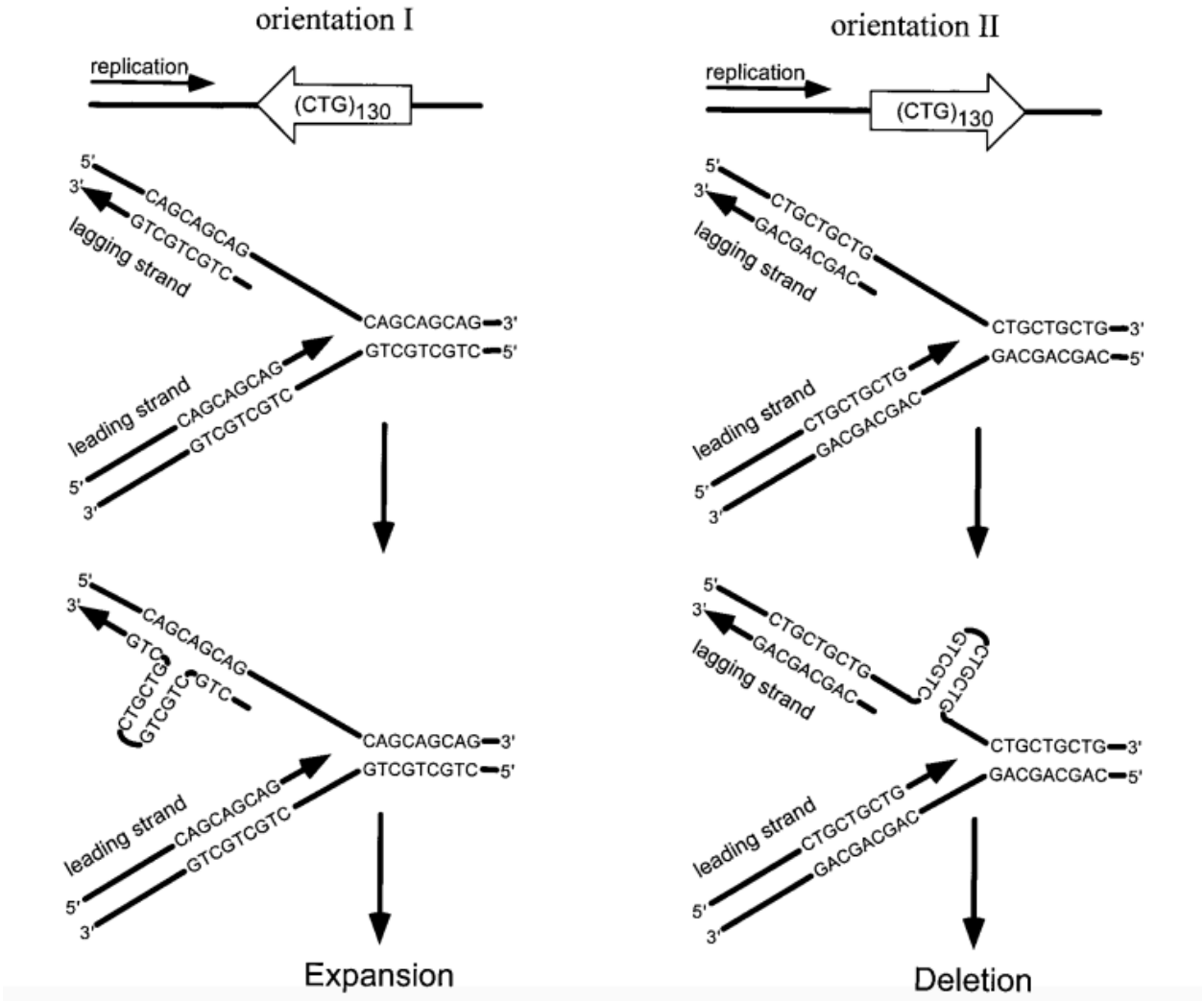


Figure 3. Orientation dependence of TNRs and their instability in the genome. Hairpins preferentially form on the lagging strand over the leading strand, and CTG hairpins are more stable than CAG hairpins. When CTG repeats are present on the lagging template, there is an increased instability leading to more contractions. When CAG repeats are on the lagging template, there is less instability, but more expansions are observed over contractions (Figure from Freudenreich et al., 1997).

The hairpins formed by both CAG and CTG repeats are stable at physiological temperatures, implying that the only way to resolve the structures is through cellular repair mechanisms (Petruska et al., 1996). Therefore, DNA repair and TNR instability are intimately

linked. Multiple pathways of both single-strand break repair and double-strand break repair have been implicated in repeat expansions across a wide array of organisms. While the initial model for repeat expansions relied on the role of replication in instability, DNA repair, particularly double-strand break repair, has recently emerged as another cause of expansions and contractions (Polleys et al., 2017). The two main models for TNR expansion are through replication and repair. It is hard to parse out when instability is due to one over the other, as TNRs are fragile and often break during replication, and therefore replication and repair often go hand in hand (Kovtun et al., 2008).

Different numbers of repeats can be involved in hairpin formation, and that in turn affects the stability of the hairpin and the ability of different DNA repair proteins to resolve the structure. One mechanism of DNA repair, mismatch repair, or MMR, is involved in resolving hairpins made up of 8 repeats in the stem of the hairpin (totaling 4 mismatches) (McMurray, 1999). While MMR can resolve some hairpins, is unable to recognize and correctly repair all. This is most likely not due to a defect in the MMR proteins, but instead due to a defect at the DNA level (i.e. DNA secondary structures) that the proteins are unable to recognize. Mutations in the MMR proteins are therefore not causing repeat expansions, as would be expected since if this were the case, there would be an increase in mutations genome-wide, and not solely in the specific locus of the TNRs (Kovtun et al., 2008). Rather, MMR proteins are likely causing expansions by acting abnormally at the triplet repeat sequence.

It mammals, it is believed that base-excision repair, or BER, is the underlying cause of many repeat expansions. BER is a repair pathway that can remove oxidative damage of DNA bases. OGG1 is the enzyme in BER that is principally responsible for CAG repeat expansions, as it removes 7,8-dihydro-8-oxoguanines (or 8-oxo-G), a type of oxidative damage in which a

reactive oxygen species alters a guanine in the DNA (which are very prevalent in GC-rich CAG TNRs). Oxidative damage is not a direct consequence of having long tracts of repeats, but instead is correlated with aging. Repair of the increased oxidative damage is implicated in repeat expansions and is noticeable in the brains of HD patients. This implies that TNR instability due to DNA repair is a good hypothesis, as neurons are terminally differentiated cells, and do not go through replication (Kovtun et al., 2007).

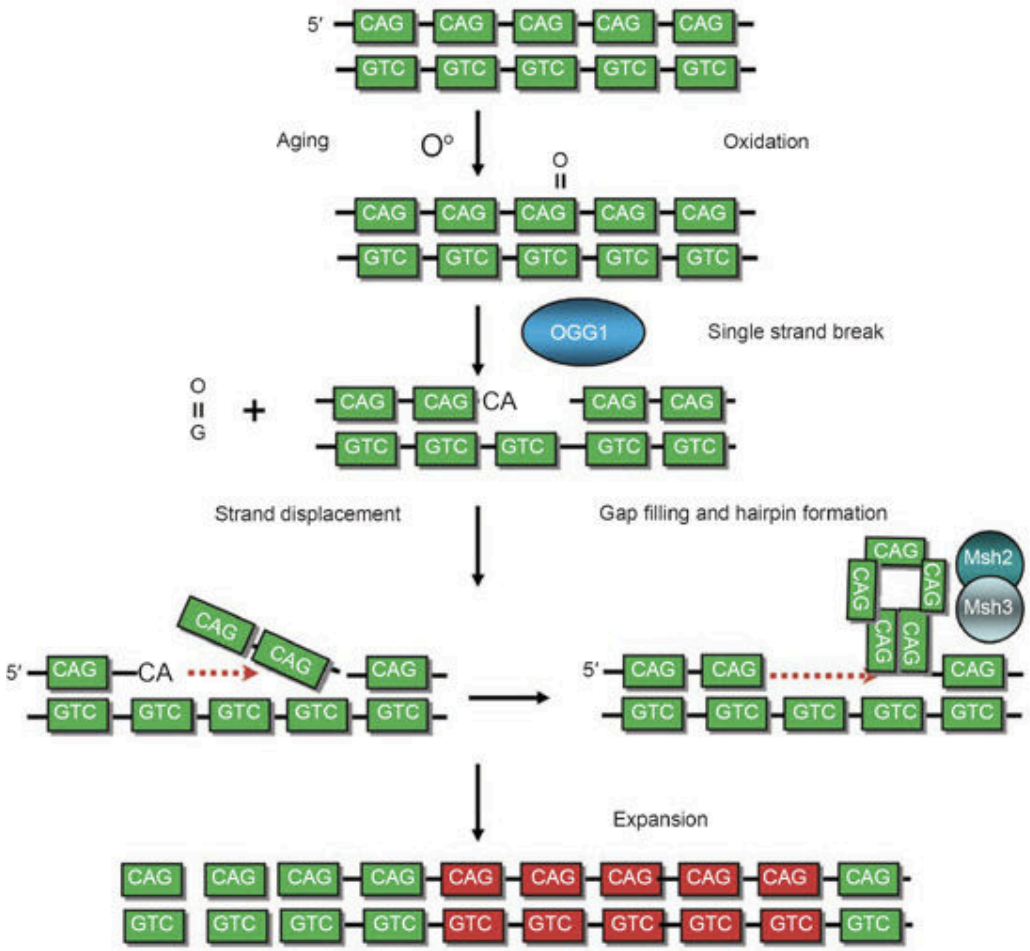


Figure 4. A model for repeat expansion via BER. Reactive oxygen species generated from cellular metabolism in the mitochondria react with the DNA to form oxidative damage that gets repaired via BER. OGG1 creates a single-strand break in the DNA in order to remove the damaged base. In unstable regions of TNRs, the break can often lead to hairpin formation and subsequent expansion. The red repeats are the added repeats in the expansion (Figure from Kovtun et al., 2008).

Additionally, Msh2 and Msh3, two repair proteins involved in MMR, have been linked to repeat expansions in mice (Kovtun et al., 2008). As shown in the figure above, Msh2/3 can bind to a hairpin due to the non-Watson-Crick base pairing and can perpetuate the hairpin structure long enough for the gap-filling synthesis during BER to create an expansion event (Kotvun et al., 2008). Msh2 can also recognize slipped-strand structures in the DNA to repair via MMR (Lenzmeier et al., 2003).

TNRs also impede the replication fork because of secondary structure formation. Different organisms show different length-dependence of repeats in order to cause replication fork stalling and pathology. One hypothesis behind this phenomenon is in the difference between GC content of the genome in different species. If an organism's genome is more AT-rich, then GC-rich repeats are more novel, and there may not be as many safe-guards against forming secondary structures (Mirkin & Mirkin, 2007).

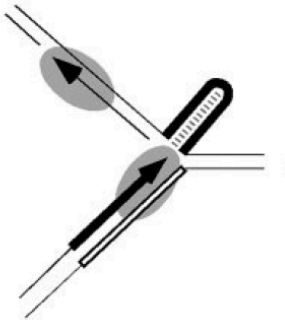


Figure 5. TNRs and fork stalling due to hairpin formation on the template. Figure adapted from Mirkin and Mirkin, 2007.

Mus201, a gene involved in transcription-coupled repair and nucleotide excision repair (XPG in humans), has been shown to play a role in TNR instability in *Drosophila*, as null mutants exhibited significantly lowered instability and inheritance of repeat length changes through both the male and female germline (Jung et al., 2007). While many different types of

DNA damage and repair pathways have been implicated in repeat instability, the specific pathways explored through our experiments involved DNA double-strand break repair.

DNA Double-strand Breaks and Their Role in Repeat Expansion

DNA double-strand breaks (DSBs) occur when both strands of the DNA double helix are broken, usually as a result of either endogenous or exogenous damage (e.g. replication fork stalling or X-rays, respectively). While DNA DSBs are often thought of in the context of damage, there are also several vital cellular processes that rely on DSBs, such as crossing over of tetrads in meiosis or V(D)J recombination that occurs during B cell development. While DSBs can be part of normal processes, they must be repaired with high fidelity, as inaccurate repair poses a hazard to the cell. DNA DSBs are a threat to genomic integrity, as aberrant repair of these breaks often leads to loss or mutation of genetic material. Imprecise repair can therefore ultimately result in mutagenesis and tumorigenesis if left unchecked (Pardo et al., 2009).

There are two distinct, broad categories of DNA DSB repair, the first being homologous recombination (HR). HR works by first resecting the 5' ends of the DNA on either side of the break to reveal 3' overhangs. Single-stranded binding proteins such as replication protein A (RPA) then bind the overhangs to stabilize them. RPA is displaced by BRCA2 (Rad52 in yeast) in order to load Rad51, which catalyzes the invasion of a sister chromatid or homolog with an identical or highly similar sequence, respectively. After invasion, the strand elongates and copies off of the homologous template sequence in order to regain any sequence that was lost in the damage, then returns to the break and anneals. While HR is lauded as the more high-fidelity mechanism of repair, as it often utilizes an identical sequence on a sister chromatid, it can still be error-prone (Rodgers et al., 2015). In addition to this, in a diploid organism, HR can use the homolog instead of the sister chromatid for repair. Because homologs come from different

parents and can contain alternate alleles for a gene, on occasion, HR at these loci can result in a phenomenon known as loss of heterozygosity (LOH), which is a necessary step in the formation of several types of cancers (Stark et al., 2003). Additionally, HR in the context of TNRs can also be highly inaccurate and result in increased instability (reviewed in Polleys et al., 2017).

Among the main types of homologous recombination, there exists synthesis-dependent strand annealing (SDSA) and classical double-strand break repair (DSBR). In SDSA, only one end of the DSB invades and copies off of the homologous template. After invasion and elongation through to the other homologous region, the strand is displaced, effectively resolving the D-loop, and the invading strand returns to the break to reanneal with the other end of the break. Because only one side of the break invaded the homolog and repaired, the other strand (from the other side of the break) then uses the newly-synthesized strand as a template for its repair, known as gap-filling. Therefore, SDSA does not result in a crossing-over event. The other type of homologous recombination, classical double-strand break repair (DSBR), invades the homologous template from both sides of the break, resulting in double Holliday Junctions (double HJs). These HJs can be resolved in order for the invading sequences to return to the break and anneal, and depending on how they get resolved, can either produce or not produce a crossing-over event (Pardo et al., 2009).

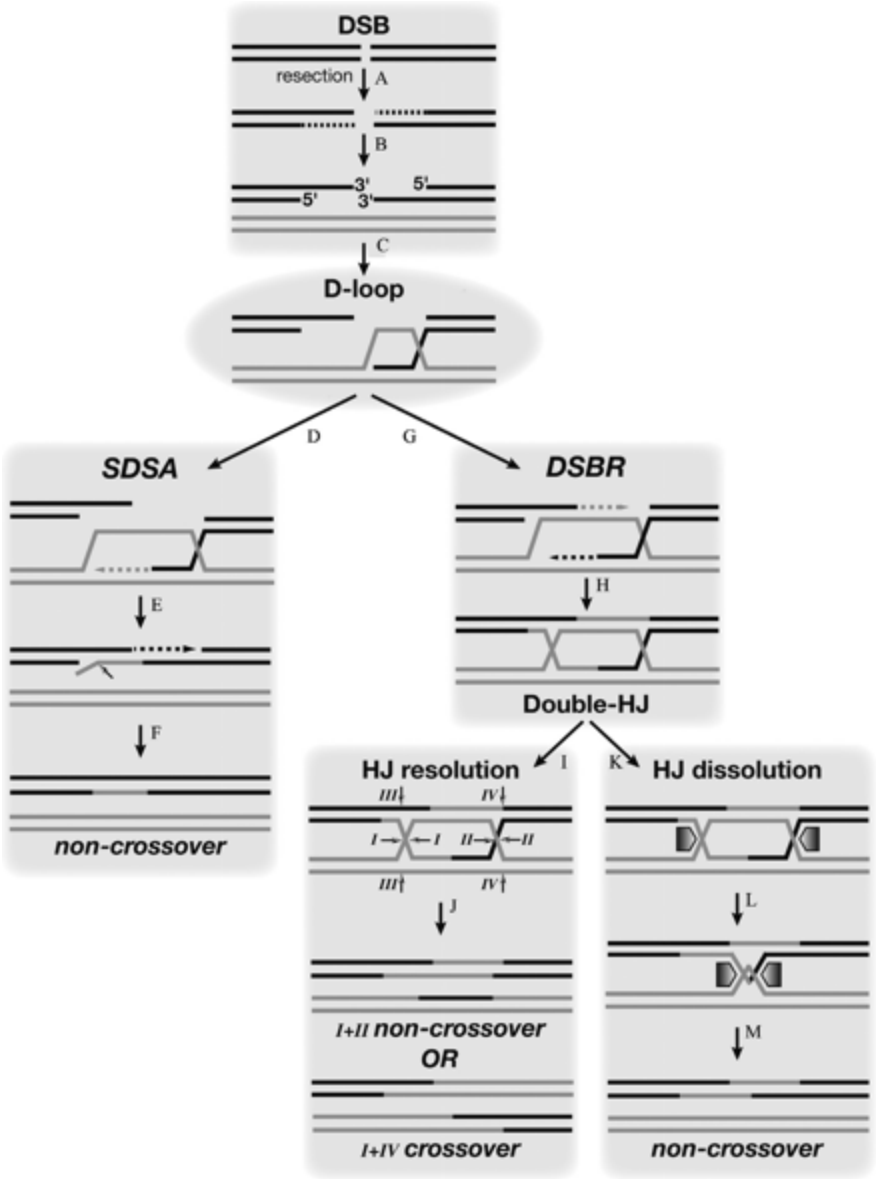


Figure 6. Types of homologous recombination. After the induction of a DNA DSB (A), if the cell goes through HR, the 5' ends of the break are first resected (B). The 3' overhang then invades a homologous chromosome or sister chromatid for repair (C). If the repair pathway choice is SDSA, only one side of the break invades and copies off of the homologous sequence (D). After strand displacement, the nascent strand returns to the break, so the other strand can use it as a template for repair (E). The resulting repair event does not create a crossover (F). If the cell repairs via DSBR (G), then both sides of the break invade to create a double HJ (H). The HJ can be resolved to either create a crossover (I-J) or a non-crossover (K-M) (Pardo et al., 2009).

The other type of DNA DSB repair, generally regarded as the more error-prone pathway, is Non-Homologous End Joining (NHEJ). This pathway does not include the use of a

homologous sequence, and instead simply ligates the two ends of the break back together after damage. Canonical end-joining utilizes the heterodimer Ku70/80, which is thought to stabilize the ends of the break, preventing resection. The ends are then ligated via ligase IV. There is no resection during C-NHEJ (or c-EJ), and often there is no synthesis, meaning any sequence lost in the damage is not re-synthesized during repair. Because of this, end joining can often result in small deletions from the genome. In vertebrates, it has been shown that polymerases lambda and mu are able to perform *de novo* synthesis, leading to non-templated insertions during repair via NHEJ (Pryor, 2015). Underneath the umbrella of NHEJ, there is a type of alternative end joining, or Alt-EJ, called Microhomology-mediated end joining (MMEJ). MMEJ does not require the Ku70/80 heterodimer, and therefore exhibits resection of the 5' ends to reveal small homologies of approximately 5-25 nt in length (Pardo et al., 2009). These microhomologies in the sequences can be annealed. Due to the mechanism of MMEJ, deletions always occur during this type of repair (McVey et al., 2008). A third type of NHEJ, single strand annealing, or SSA, also forgoes the use of Ku70/80 in order to resect the 5' ends of the break, this time to reveal longer homologies on the 3' overhangs. Similar to MMEJ, the homologies then anneal and the non-homologous overhangs, or flaps, are trimmed by an endonuclease, and the gaps are filled in via synthesis (McVey et al., 2008) (See figure 7). Many consider SSA a type of HR instead of NHEJ, as it utilizes Rad52 (an ortholog of BRCA2) to anneal the two strands of DNA, although it does not involve invasion or repair off of a homologous sequence (Sugawara, 2000).

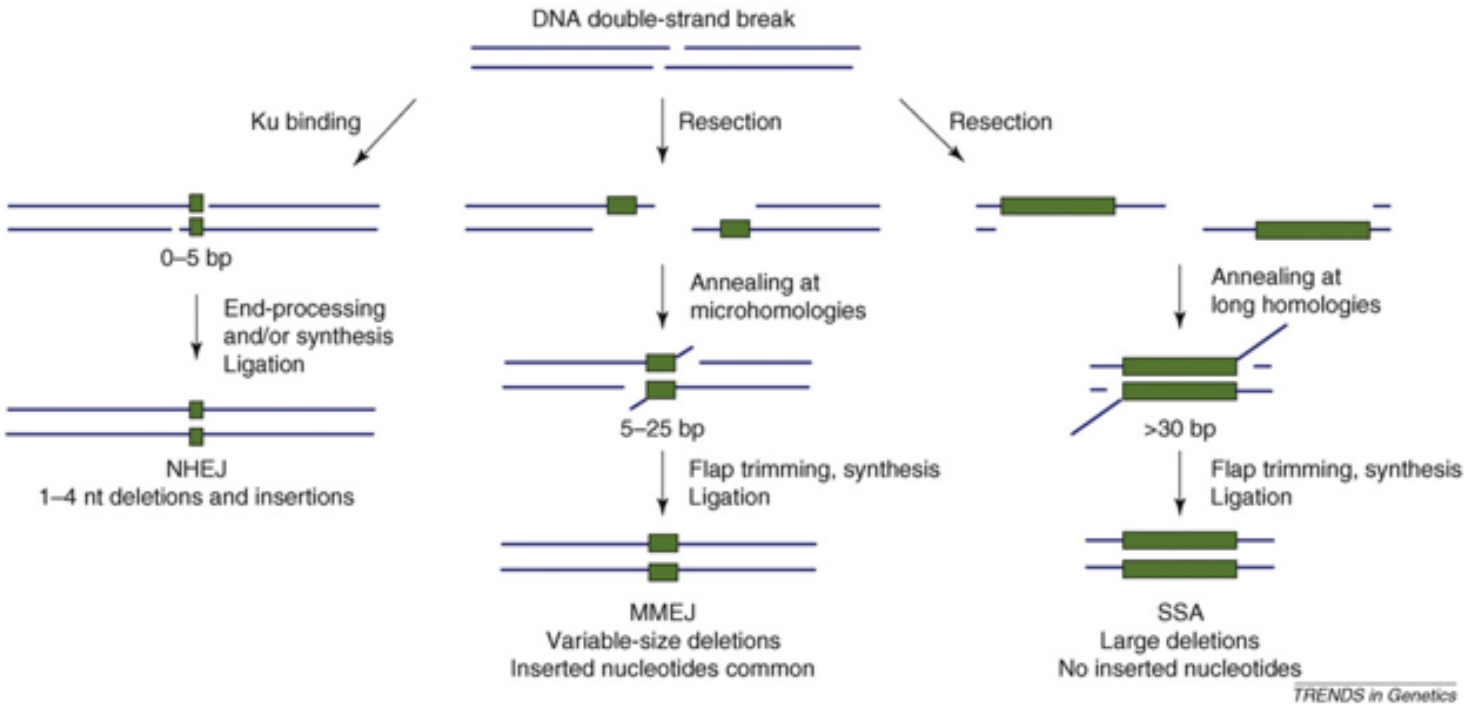


Figure 7. Types of end joining. Far left pathway shows canonical non-homologous end joining, where there is end tethering by the Ku70/80 heterodimer and minimal resection. Middle pathway models microhomology-mediated end joining with resection of the 5' ends to reveal microhomologies on the 3' overhangs that ultimately anneal and the flaps are clipped via an endonuclease. The gaps are then filled-in via synthesis. Far right pathway demonstrates single strand annealing, where resection of the 5' ends reveals larger homologies on the 3' overhands that can bind, resulting in larger deletions than in MMEJ. (McVey et al., 2008)

TNR instability has been implicated in DNA DSB repair through the roles of multiple proteins, including the MR(X)N complex, which is involved in both HR and NHEJ (Richard et al., 2000). This same study also found that in yeast, upon the induction of a DSB, recombination repair led to more instability (including both expansions and contractions) than is observed through replication. Similar to replication-induced instability, however, it is also believed that secondary structure formation during synthesis is the main cause of TNR instability during recombination repair (Richard et al., 2000). Specifically, homologous recombination after the induction of a double-strand break has been shown to lead to a higher amount of instability, increasing the frequency of both expansions and contractions (13% expansion events as

compared to 0% in the no-break control, and 30% contractions as compared to 10% in the control) (Richard et al., 2000). Additional experiments in yeast have shown that during repair via homologous recombination through a region of tandem repeats, expansion and contraction events were recovered approximately 50% of the time on the recipient molecule after repair (Pâques et al., 1998). The results from this paper further supported the theory that the copying of a template with repeats during repair was causing the instability, similar to the mechanism that is observed during replication-induced instability of repeats.

In another yeast study, secondary structure-prone sequences or repeats were inserted into the genome, and replication slippage was observed. These slippage events were dependent on RAD52 and RAD50, which were responsible for up to 90% of the contractions observed in the study (Tran et al., 1995). Rad52 helps load Rad51 onto RPA-coated ssDNA and also plays a role in annealing the DNA in second-end capture during SDSA (Nimonkar et al., 2009). *Drosophila* do not have Rad52, but BRCA2 is the closest homolog (Liu et al., 2011). Events of replication slippage that do not get repaired lead to the instability of these repeats (Kunkel, 1993).

Goals of This Study

As there are multiple studies confirming the instability of trinucleotide repeats in yeast (a single-celled organism) after DNA double-strand break repair via homologous recombination, it was of specific interest to elucidate if the same phenomenon could be observed in a multi-cellular system, such as the fruit fly. Additionally, owing to the plethora of evidence in support of germline expansions causing genetic anticipation, *Drosophila* served as a model organism in which to study the tissue-specific instability of trinucleotide repeats. The dependence of some models of repair-induced repeat instability on RAD52 (BRCA2) implied that several pathways of repair could be involved, including SDSA. The specific goal of this study was to further

elucidate the role of DNA double-strand break repair, particularly the Rad52-dependent SDSA pathway, in the instability of CAG/CTG trinucleotide repeats in the metazoan system of *Drosophila melanogaster*.

Materials and Methods

Fly Crosses

In order to get the DNA repair assay constructs of the donor and recipient chromosomes into a stock of flies, the constructs first had to be molecularly cloned (see Molecular Cloning section of Materials and Methods). The final plasmids were midiprepped following the Macherey-Nagel Plasmid DNA Purification (NucleoBond Xtra Midi) protocol (Macherey-Nagel, 2014). Following midipreparation and quantification of DNA concentration via Nanodrop spectrophotometry, the plasmids were shipped to BestGene Inc. to be injected into *Drosophila* embryos. The constructs were integrated into the genome (on the 2nd chromosome at cytological site 59D3) via PhiC31 integrase. PhiC31 integrates the injected plasmid via its attB site (see figure 13) with the attP docking site in the genome of the injected flies. The surviving flies from the injected embryos were then crossed to yw flies (flies with a brown body and yellow eyes) in order to remove the PhiC31 source in the next generation. The red-eyed progeny of that cross were selected and shipped back to the McVey Lab. The red-eyed progeny were selected, as they contained the integrated plasmid, because the attB site is marked with the White gene, which codes for red eyes. These flies contained the integrated plasmid on one of the homologs for chromosome II, the other homolog being wild type. After receiving successful transgenic flies for both the recipient and donor constructs, these stocks were crossed to create the males for the final single male cross, which contained the donor and recipient chromosomes, as well as *I-SceI* (see supplemental figure 1 for fly cross scheme).

Molecular Cloning

pattB-H1-LacZ-H2 plasmid (hereafter referred to as pSD028) was provided by Sarah Dykstra (see figure 13). pSD028 served as the backbone for all further cloning of the donor chromosome for the assay. From pSD028, a multiple cloning site containing the cut sites needed for cloning eGFP and the TNRs along with backup cut sites was cloned in to create pJB026.

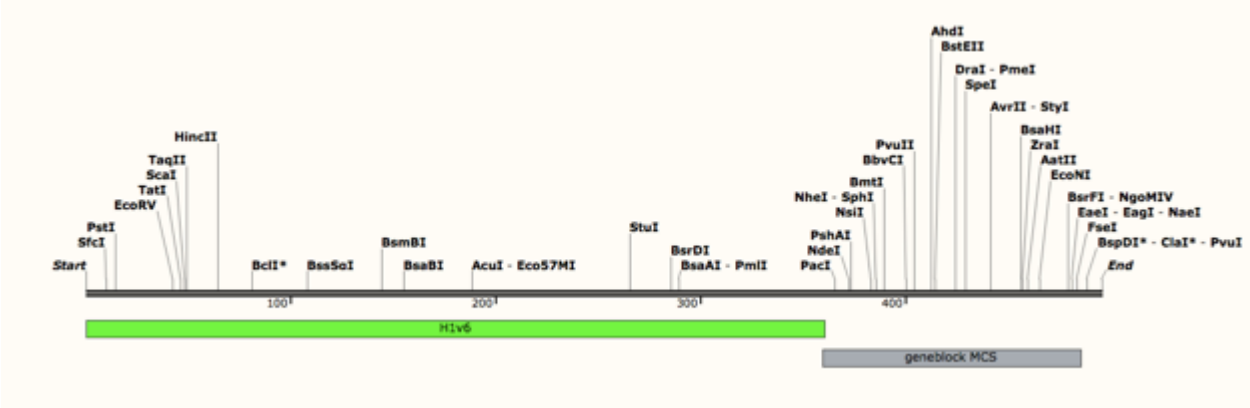


Figure 8. MCS containing cut sites for the enzymes used to clone in eGFP and TNRs. Backup cut sites also included in case cloning scheme needed to be changed or enzymes were not cutting effectively. The gene block also contained the portion of the first homology arm that was lost in pSD028 when it was digested for cloning.

The multiple cloning site was created as a gene block by Integrated DNA Technologies (IDT) and was shipped back to the McVey lab contained in a plasmid (hereafter referred to as pSD025). The MCS and portion of homology arm 1 from pSD025 were directionally cloned into pSD028 by digesting both pSD028 and pSD025 with FseI and PstI-HF to create ligatable sticky ends. PstI was located within the first homology arm of pSD028, meaning the gene block had to be engineered to contain the lost portion of the homology arm, in order to ensure more efficient repair via HR during the assay, as well as the MCS for future rounds of cloning.

1.5 µL pSD028 was digested using 1 µL each of restriction enzymes FseI and PstI-HF, along with 5µL CutSmart buffer, and 41.5 µL ppH_2O. The high-fidelity version of PstI was used as it was compatible in CutSmart buffer. The 50 µL total reaction volume was digested overnight

(16 hours) at 37 °C. The digestion of pSD028 with FseI and PstI-HF excised a 3645 bp fragment that included all of LacZ and part of the first homology arm. The remaining 8772 bp backbone that was used for cloning contained the remainder of the first homology arm, along with the second homology arm, the white gene, the ampicillin resistance gene, and the attB integration sequence. A second 37 °C overnight digest was set up of 1.5 µL pSD025 (see figure 8) in order to excise the insert, which was 495 bp. Again, 1 µL of each FseI and PstI-HF was used alongside 41.5 µL ppH_2O and 5 µL CutSmart buffer.

The overnight reaction mixtures were then electrophoresed at 100 V and the products of size corresponding to the digested vector (approximately 8.8 kb) and digested insert (approximately 470 bp) were excised and purified following the Nucleospin Gel and PCR Clean-Up Kit protocol 5.2 from Macherey-Nagel (Vogelstein et al., 1979). The concentrations of DNA were then measured via NanoDrop Spectrophotometry.

The digested vector and insert were then ligated in a 1:3 and 1:5 vector to insert ratio. XµL vector, 1µL T4 10X Ligation Buffer (NEB), 1µL annealed oligo insert, 1µL T4 DNA Ligase (NEB), 10µL-XµL pico-pure H_2O, then placed at 25 °C for 1 hour. The two ligations were performed in duplicate, with one serving as the control without any oligo insert added to the reaction. The duplicate reaction controlled for any vector that may have re-annealed and provided a level of background for the transformation.

XL10 chemically competent *E. coli* cells were used for the transformation. These cells were created following the Inoue protocol (Sambrook et al., 2006). Chemically competent cells were thawed on ice for approximately 10 minutes before 3µL DNA (75-100ng) was added. The added DNA consisted of either vector + insert, no insert controls (vector only), or pUC19 (where only 1 µL was used) to control for transformation efficiency. The cells were placed on ice for 30

minutes, then heat shocked for 30-45 seconds at 42 °C to allow the cell membranes to become more permeable to the ligated plasmid DNA. The cells were then immediately moved back to the ice for 1 minute, then rescued in 200 μL LB for one hour at 37 °C on a rotator. The cells were plated on LB + Amp plates and grown overnight at 37 °C.

After incubating at 37°C for approximately 16 hours, single colonies from the vector + insert plates were inoculated into 2 mL LB + 2 μL Amp and spun on a rotator at 37 °C for 8 hours, then miniprepped following NucleoSpin Plasmid EasyPure protocol from Macherey-Nagel (Macherey-Nagel, 2014). A diagnostic digest of the miniprepped single colonies was then set up using the same enzymes used for cloning (FseI and PstI-HF in this case) and put at 37 °C overnight. The overnight diagnostic digests were then electrophoresed at 100 V to confirm presence of the insert in the plasmid.

After confirmation of the insert, the plasmid was midi prepped following the Macherey-Nagel Plasmid DNA Purification (NucleoBond Xtra Midi) protocol (Macherey-Nagel, 2014). The concentration of DNA in the MidiPrep was determined using a NanoDrop spectrophotometer. The midi prep was then sent to Eton Bioscience for sanger sequencing. Presence of the correct homology arms, as well as the MCS was confirmed via sequencing.

Table 1. Primers used to sequence plasmids prior to injection.

Primer Name	**Primer Sequence (5'-3')**	**Primer Use**
newCAGFor	CCTCAGCCTGGCCGAAAGAAAGAAA	Forward primer through CAG repeats
T720B	TAATACGACTCACTATAGGG	Reverse primer through CAG repeats

SD101	GGCATGTCGACACTAGCGGATC	Forward primer through homology arm 1
SD102	TTTCTTTCTTTCGGCCAGGCTGAGG	Reverse primer through homology arm 1
SD0093	CACTGCATTCTAGTTGTGGTTTGTCC	Forward Primer through homology arm 2
pUC57 seq 5932 R	GCGGCCTTTTTACGGTTCCTG	Reverse primer through homology arm 2
SD094	CGGGTCACCGTTTAAACACTAGTATG	Forward primer through eGFP, starting at terminator
SD106	CGCCTAGCTAAGGGACGTCG	Reverse primer through eGFP starting at SV40 promoter
SD030	GACATCCAGTGTTTGTTCCTTGTG	Forward primer through attB
SD059	GTCATAGCACTAGACCAGTCGAC	Reverse primer through attB
SD067	GACTGGTCCAGTTCACAACGT	Forward primer through MCS/H1 gene block
SD049	GCAAGTAAGGAACATAGCATACCACCCG	Reverse primer through MCS/H1 gene block

After the confirmation of the insert via sequencing, the newly created plasmid containing the MCS (hereafter referred to as pJB026), was used to create the following plasmids for injection (See figure 13-30 in results):

Table 2. Cloned plasmids. The plasmid backbone and insert used to create the resulting plasmid listed along with the enzymes used to cut both backbone and insert.

Backbone Plasmid	**Insert**	**Enzymes Used**	**Resulting Plasmid**
pSD028	pSD025 (MCS + part of H1)	PstI and FseI	pJB026
pJB026	Plasmid 1285 (containing eGFP) from the DGRC)	AatII and AvrII	pJB027
pJB027	pCF390 (CAG70 repeats)	BstEII and either PshAI (pJB027) or PvuII (pCF390)	pJB028
pJB027	pCF391 (CAG130 repeats with two templated insertions)	BstEII and either PshAI (pJB027) or PvuII (pCF391)	pJB029
pJB026	pCF390 (CAG70 repeats)	BstEII and either PshAI (pJB026) or PvuII (pCF390)	pJB032
pJB026	pCF391 (CAG130 repeats with two templated insertions)	BstEII and either PshAI (pJB026) or PvuII (pCF391)	pJB033
pJB027	pCF590 (CTG70 repeats)	BstEII and either PshAI (pJB027) or PvuII (pCF590)	pJB030
pJB026	pCF590 (CTG70 repeats)	BstEII and either PshAI (pJB026) or PvuII (pCF590)	pJB034

The same chemically competent cloning protocols as stated above were used to clone the rest of the plasmids, and only the enzymes used for cloning were changed. After confirmation of successful clones via Sanger sequencing, the plasmids were frozen down as glycerol stocks: a single colony containing a successful clone was inoculated into 2 mL LB + Amp overnight until the culture reached log phase. 500 μL culture was then added to a cryovial, and then 500 μL of 10% glycerol was also added. The resultant glycerol stock was inverted several times in order to homogenize the mixture before storing at -80 °C.

Scoring of the Assay

Flies were scored according to the type of repair event on the recipient chromosome they inherited. Progeny inherited either the donor or recipient chromosome from males constitutively expressing *I-SceI* on the X chromosome (position 5B with the ubiquitin promoter (characterized by Preston et al., 2006, see supplemental Table 2)) and both the donor and recipient 2nd chromosome homologs. Every cell of the males of that generation, including the germ line cells, ideally should have had a unique cutting and repair event occurring. The repair events from the germ line could then be passed on to the next generation (see supplemental figure 1 for fly cross) and categorized according to the type of fluorescence seen in the progeny that inherited the recipient chromosome. The recipient chromosome was also marked with sternal plural (Sp), a dominant phenotypic marker. This allowed the progeny that inherited the recipient chromosome to be scored first by the presence of Sp, then by the type of fluorescence. All male flies of that final generation were collected, including the non-Sp males (males that had received the donor chromosome) in order to test for CAG repeat length on the homolog where a repair event had not occurred. The Sp possessing flies were then further scored by fluorescence: DsRed+/GFP-

(indicating a type of end-joining event, most likely C-NHEJ), DsRed-/GFP- (indicating MMEJ or aborted SDSA), or DsRed-/GFP+ (indicating complete SDSA).

The flies were scored under a fluorescent microscope, using the filter for GFP2 or DsRed to score for the presence of either GFP or DsRed, respectively. The DsRed+/GFP- flies were batch collected in an Eppendorf tube according to their isolate number and stored at -20°C. Similarly, all non-Sp flies were collected and stored *en masse* this way. Any DsRed-/GFP+ or DsRed-/GFP- Sp males were collected in 96 well plates and stored at -20°C. The 96 well plates had columns 1-12 and rows A-H. Each row corresponded to a single isolate (all the progeny from one isolate were from the same single male cross), and individual progeny were collected across the columns of their specific isolate row.

	1	2	3	4	5	6	7	8	9	10	11	12
A												
B												
C												
D												
E												
F												
G												
H												

Figure 9. Example of a 96 well plate.

The first 8 wells of each row were dedicated to DsRed-/GFP- progeny, while the final four wells were for DsRed-/GFP+ progeny. If there were not enough progeny to fill the wells, the wells were left empty. If there were too many progeny of a specific phenotype to fit in the allotted wells, the extra flies were batch collected in Eppendorf tubes according to isolate number. Both the Eppendorf tubes and 96 well plates were stored at -20°C.

Using the camera on the fluorescent scope, pictures were taken on several of each different repair event. Conditions for each were as follows:

Bright field pictures taken with a 500 millisecond maximum exposure time with a gain of 1, DsRed filter pictures taken with a 500 millisecond maximum exposure time with a gain of 4, and GFP2 filter pictures taken with 300 millisecond maximum exposure time with a gain of 4 (see figures 34 and 35).

Synthesis Tract Length PCR

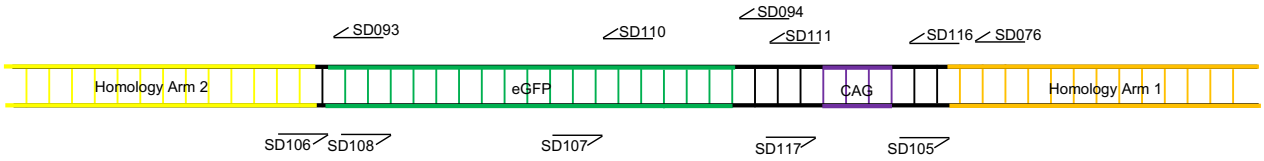


Figure 10. Primers used for synthesis tract length PCRs in final generation males. (see table 3)

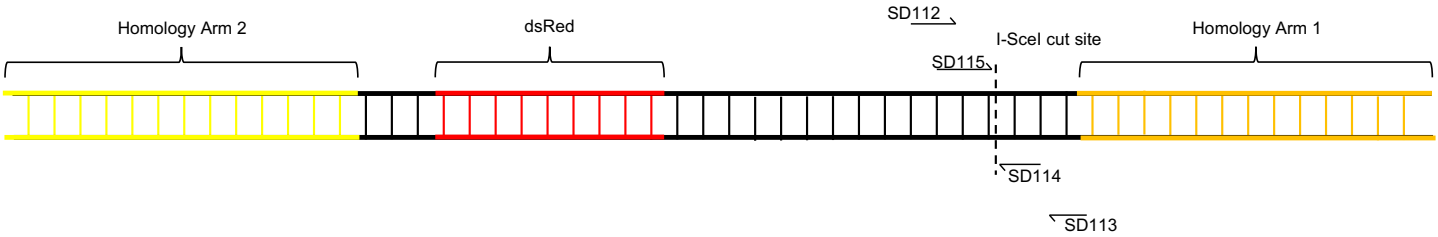


Figure 11. Primers used for end joining PCRs in final generation males. (see table 4)

The DNA from the collected flies was prepared for PCR via fly prep: the frozen down flies (stored at -20°C) were first squished with 49 μL squishing buffer (10 mM Tris-Cl at pH 8.2, 1mM EDTA, and 25 mM NaCl), then 1 μL proteinase K was added. The flies were then put in a thermal cycler to be kept at room temperature (25°C) for 30 minutes followed by 2 minutes at 95°C to heat inactivate the proteinase K. The fly prep could then be stably stored at either 4°C or -20°C. The samples could then be run in a touchdown PCR with the primers from Table 3 to determine the type of repair and the approximate size of the repeats in the collected fourth generation DsRed-/GFP+ Sp males. The samples from the DsRed-/GFP- Sp males could be run in a TD PCR with the primers in Table 4 to determine the type of end joining event that could have occurred. The PCR reaction master mix, per reaction, used: 16μL ppH_2O, 2 μL 10X buffer, 0.5 μL dNTPs, 0.2 μL primer #1, 0.2μL primer #2, 0.1 μL Taq polymerase, and 1 μL DNA from the fly prep. TD thermal cycler conditions for determining CAG repeat length: initial denaturation at 95°C for four minutes, followed by 16 cycles of 95°C denaturation for 30 seconds, 62°C annealing for one minute (stepping down at -0.5°C per cycle), and 68°C extension for three minutes. The first 16 cycles were followed by another 20 cycles, with an annealing time of 60 seconds at 55°C (with no stepdown). Lastly, there was a final extension at 68°C for seven minutes before an indefinite hold at 4°C. Correct conditions have yet to be determined for synthesis tract length PCR or end joining characterization PCR.

Table 3. Primers used to determine synthesis tract lengths.

Primer Name	Primer Sequence (5'-3')	Primer Use
SD076	GACGCACTCACTAACGATGATGAG	Forward primer to sequence across whole synthesis tract length assuming invasion from H1
SD093	CACTGCATTCTAGTTGTGGTTTGTCC	Forward primer to be used in conjunction with SD106 to see if invasion occurred from H2
SD094	CGGGTCACCGTTTAAACACTAGTATG	Forward primer to sequence through GFP
SD110	CCATCCTGGTCGAGCTGGAC	Forward primer in GFP, approximately 1 kb of synthesis away from H2
SD111	GCCGCGGTGGAGCTCGAATTC	Forward primer before GFP, approximately 1.5 kb of synthesis away from H2
SD116	AGTTTGCCCATCCAGGTCAG	Forward primer in between H1 and CAGs to sequence through repeats
SD105	CAAACTGGAGGCCTGGGAAG	Reverse primer outside of H1 to see if invasion occurred from H1
SD106	CGCCTAGCTAAGGGACGTCG	Reverse primer to sequence across whole synthesis tract length assuming complete homologous recombination (or invasion from H2)
SD107	CGCTCCTGGACGTAGCCTTC	Reverse primer in GFP, approximately 1 kb of synthesis away from H1
SD108	TTATGATCTAGAGTCGCGGC	Reverse primer in GFP, approximately 1.5 kb of synthesis away from H1
SD117	GCTCGAAGGGTCCTTGTAGC	Reverse primer outside of CAGs to sequence through repeats

The PCR products were then electrophoresed on a 1% agarose gel at 100V. PCR products of the TNRs electrophoresed on a 2% agarose gel at 70V, as a higher percentage agarose and lower voltage was able to more precisely separate the repeats, which are prone to forming secondary structures. Presence or lack of a PCR product, as well as the size of the product indicated what type of repair occurred. The CAG repeat size could only be accurately resolved by approximately three repeats on a 2% agarose gel, and therefore a fragment analyzer (able to resolve repeat length by one repeat difference) was used to determine repeat length in the flies.

Table 4. Primers used to determine type of end joining.

Primer Name	**Primer Sequence (5'-3')**	**Primer Use**
SD112	CTCGAGGCCTCGAGTTAACG	Forward primer to show MMEJ in recipient
SD113	GACTGTGCGTTAGGTCCTGTTC	Reverse primer to show MMEJ in recipient
SD114	CGTTTAGAGCAGCAGCCGAATTC	Reverse primer to show c-NHEJ in recipient
SD115	GATCCACTAGTGGCCTATGCG	Forward primer to show c-NHEJ in recipient

The primers used to determine type of end joining (c-NHEJ vs. MMEJ) were strategically placed at certain distances away from the *I-SceI* cs. SD112 and SD113, used to check for MMEJ, were placed either 68 bp upstream or 88 bp downstream of the cut site, respectively. SD114 and SD115, used to check for c-NHEJ, were placed either 14 bp upstream or downstream of the cut site, respectively. Because MMEJ resects the ends of the break, the primers used for c-NHEJ would have been deleted, and therefore no product would be formed using these primers. As for the MMEJ primers, a product would be present regardless of the type of repair, but the product would run faster (representing a smaller product) for MMEJ events than c-NHEJ events.

Results

Creating the Assay

The primary goal of these experiments was to characterize any instability of the CAG/CTG trinucleotide repeats in *Drosophila* after repair of a DNA DSB via SDSA through the region of repeats. A novel DNA double-strand break repair assay was created and then the percentage of each repair event resulting from the assay was characterized. The assay utilized the second chromosome in the *Drosophila* genome and induced a DSB on one homolog using *I-SceI*, an endonuclease with an 18 bp recognition sequence, creating 5' overhangs of 4 bp in length (Monteilhet et al., 1990).

Table 5. *I-SceI* and its recognition site. Recognition site of 18 bp. Arrows indicate where the enzyme induces a break on each strand in order to produce 5' overhangs.

Enzyme	**Recognition sequence**
I-SceI	∨ 5' TAGGGATAACAGGGTAAT 3' 3' ATCCCTATTGTCCCATTA 5' ∧

After induction of the double-strand break, the DNA could repair via homologous recombination, canonical end-joining, or microhomology-mediated end-joining. The assay was biased to repair via homologous recombination by placing two homology arms, each one kilobase in length, on both homologs. In between the two regions of homology, the first chromosome, also known as the "recipient", contained the *I-SceI* cut site (*I-SceI* cs). *I-SceI* cs was located 285 bp from the first homology arm and 1873 bp from the second homology arm, thus favoring invasion for HR from the first homology arm. The recipient chromosome also contained the gene for DsRed. The other chromosome, also known as the "donor", donated its sequence to the recipient during repair via HR. The donor contained both the CAG TNRs (located 96 bp from homology arm 1) and eGFP (192 bp from the end of the CAGs and 35 bp

from homology arm 2). Total length between the two homology arms on both chromosomes was around 2 kb.

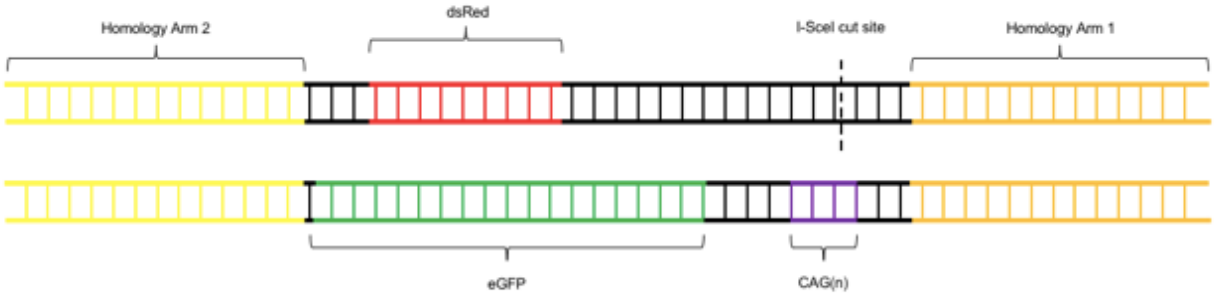


Figure 12. Model of DSB repair assay. The top chromosome is the "recipient" containing DsRed and the *I-SceI* cut site alongside of the two homology arms. The bottom chromosome is the "donor" containing eGFP and the CAG/CTG TNRs alongside of the two homology arms. Figure drawn to scale.

The creation of a novel double-strand break repair assay involved the creation of several different plasmid constructs for injection. These constructs were created in order to test this assay under different conditions (see table 2), such as with a larger number of repeats, with the repeats in different orientations, and in the presence or absence of GFP. The starting plasmid, pSD028 was provided by Sarah Dykstra and contained the two homology arms flanking LacZ, along with the White gene, attB integration sequence, and the Amp^R gene.

Figure 13. pSD028 labeled with all its elements. pSD028 contained the White gene and the attB sequence for PhiC31 integration at the attP locus. Homology arms 1 and 2 flank the sequence for LacZ (later removed). Amp^R gene present for the positive selection of the plasmid in ampicillin-containing media. pSD028 was 12,377 bp in length.

From pSD028, two parent plasmids, pJB026 and pJB027 were used to create the six final plasmids to be injected. The first, pJB026 contained both homology arms, as well as the White gene and attB integration sequence in its backbone, and replaced lacZ with a multiple cloning site (MCS).

Figure 14. pJB026. pSD028 plus MCS. Plasmid containing homology arms 1 and 2, MCS, attB, and White gene. Amp^R also present for positive selection in ampicillin-containing media. Total plasmid size of 9201 bp.

The second parent plasmid, pJB027, was created from pJB026 by cloning the sequence for eGFP into the MCS. The marker eGFP was specifically chosen for several reasons, the first of which being that it is fluorescent. The expression of a fluorescent marker creates an easily discernable phenotype that was used in the scoring of the assay. Secondly, the specific type of GFP used, eGFP, is an enhanced form of GFP that should fluoresce brighter than GFP, as it has an inserted valine after the start codon to enhance translation, and a serine to threonine mutation at position 65 in the gene (Fu et al., 2015). Because any events that would be scored in the final generation would be heterozygous for GFP, it was important to pick a fluorescent marker that would be discernable even when only one copy was present.

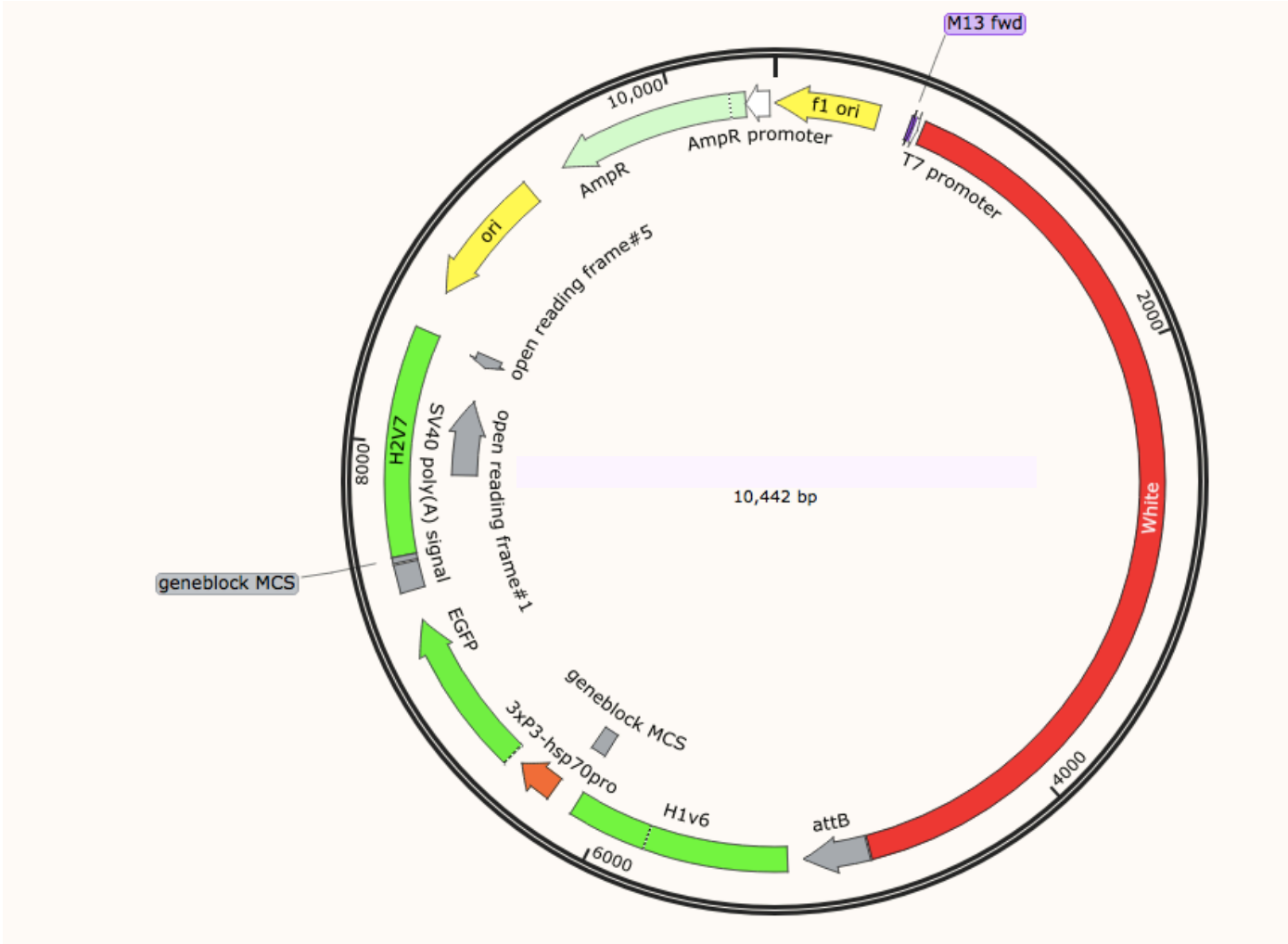


Figure 15. pJB027. pJB026 plus eGFP. Plasmid containing homology arms 1 and 2, MCS, eGFP (with promoter and terminator), attB, and White gene. Amp^R also present for positive selection in ampicillin-containing media. Total plasmid size of 10442 bp.

The sequence of eGFP was purposefully cloned in with the N-terminal region of the sequence closest to the first homology arm. The purpose of this was for twofold. Firstly, transcription would then be moving away from the CAG repeats once they were cloned in, thereby reducing the possibility of transcriptome/replisome collisions within the repeats, which would increase the fragility in that area. The second reason being that if the TNRs caused increased mutagenesis in the surrounding genes, there are more residues that, if mutated, cause loss-of-function in the protein clustered on the N-terminus than in any other part of the protein (Fu et al., 2015). Once pJB026 and pJB027 were successfully cloned, all other plasmids used for

the assay could be made. CAG or CTG repeats of length 70 or 130 were cloned into both plasmids to study repeat expansion and types of repair with and without the presence of eGFP.

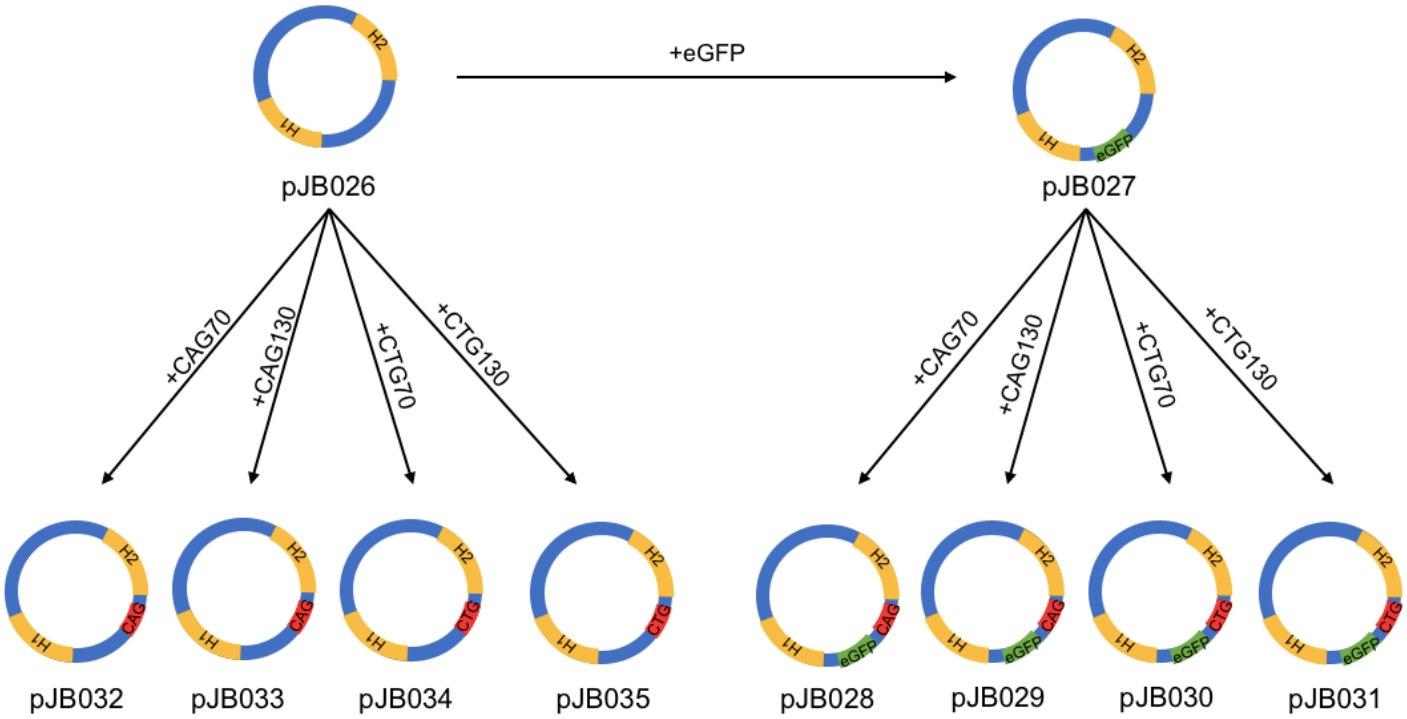


Figure 16. All combinations of donor chromosome plasmids for assay. Parent plasmid pJB026 contained the White gene and attB for PhiC31 integration, as well as the two homology arms. pJB026 was used to clone the other parent plasmid, pJB027, which also contained eGFP. From these two parent plasmids, either 70 or 130 repeats of either the CAG or CTG repeats were cloned in. Portions of the plasmids in blue represent the backbone of the plasmid, yellow represent the two homology arms, red represents the TNRs, and green represents eGFP.

In order to create the assay, a series of molecular cloning steps were done, starting with pSD028 (see figure 13). pJB026 was created from the pSD028 backbone and an MCS insert from pSD025 was successful.

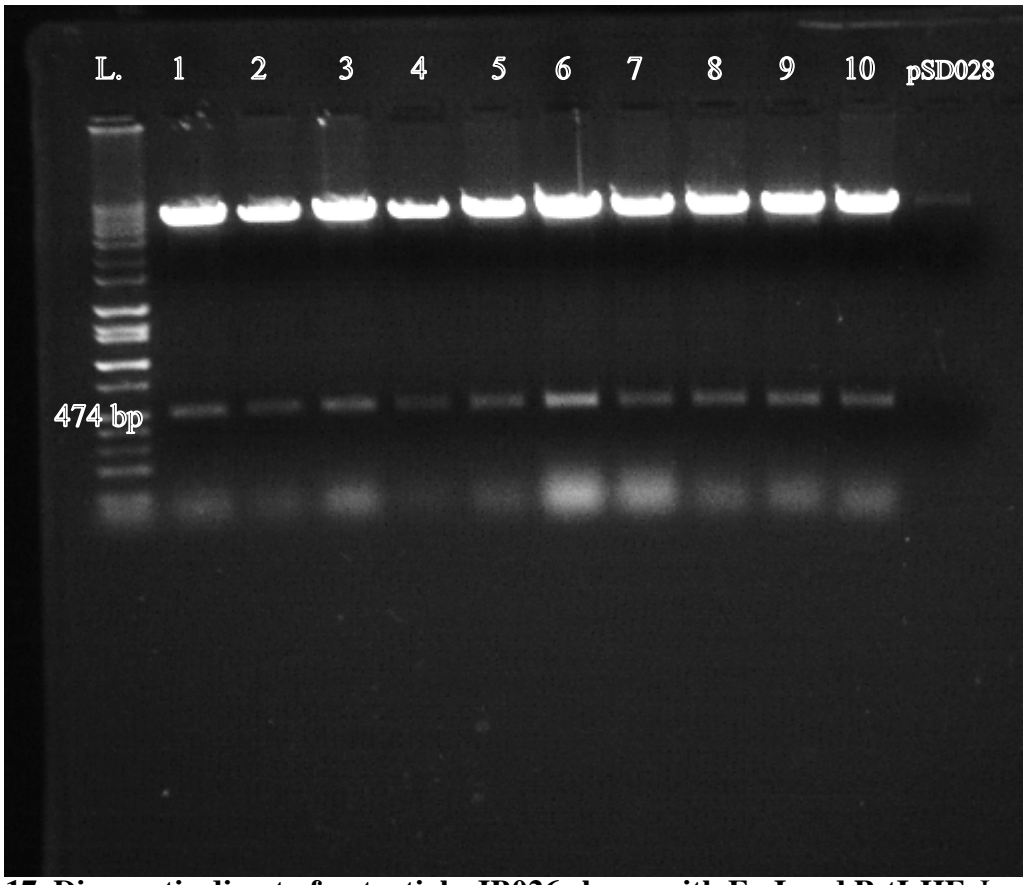


Figure 17. Diagnostic digest of potential pJB026 clones with FseI and PstI-HF. L represents the ladder. Numbers 1-10 represent the 10 colonies chosen from the plates after transformation. Products present at 474 bp indicates excised insert and confirms successful cloning of the MCS from pSD025 into pSD028 to create pJB026 in all ten colonies. Band present at top of ladder represents plasmid backbone, 8.7 kb in length.

The above gel suggests the successful creation of pJB026 by diagnostic digest. Sanger sequencing of selected clones confirmed the presence of the MCS insert (see table 1 for primers used for confirmation all molecular cloning steps via sequencing). pJB027, pJB032, pJB033, pJB034, and pJB035 could all be created from pJB026 (see figure 16 for all cloned plasmids). First, eGFP was cloned into the pJB026 backbone to create pJB027. Before creating pJB027, the correct cut sites for cloning (AatII and AvrII) were added onto the sequence for eGFP via PCR amplification with the Q5 polymerase.

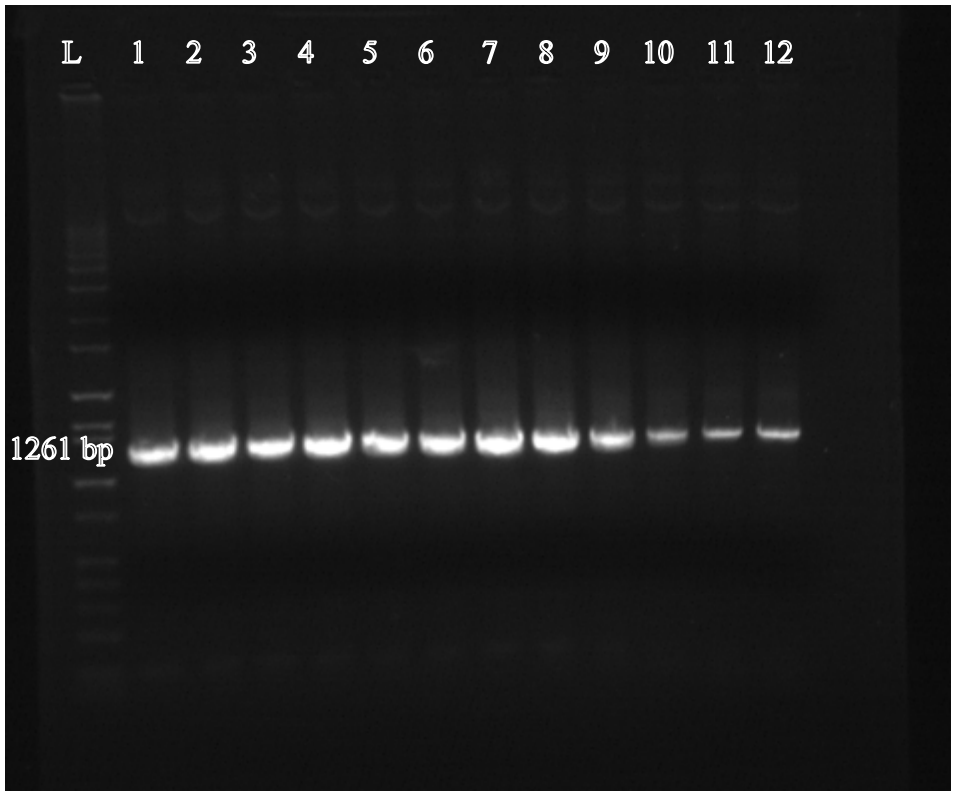


Figure 18. Gel of Q5 gradient PCR for eGFP. L represents ladder, Numbers 1-12 represent the twelve reactions in the gradient PCR reaction, with one being the reaction at the lowest Tm, and twelve being the reaction at the highest Tm. Bands present at 1261 bp confirms that the PCR worked at all melting temperatures.

The products from the Q5 PCR were then purified, digested, and inserted (following the methods previously described in the Molecular Cloning section of Materials and Methods) into the pJB026 backbone to create pJB027.

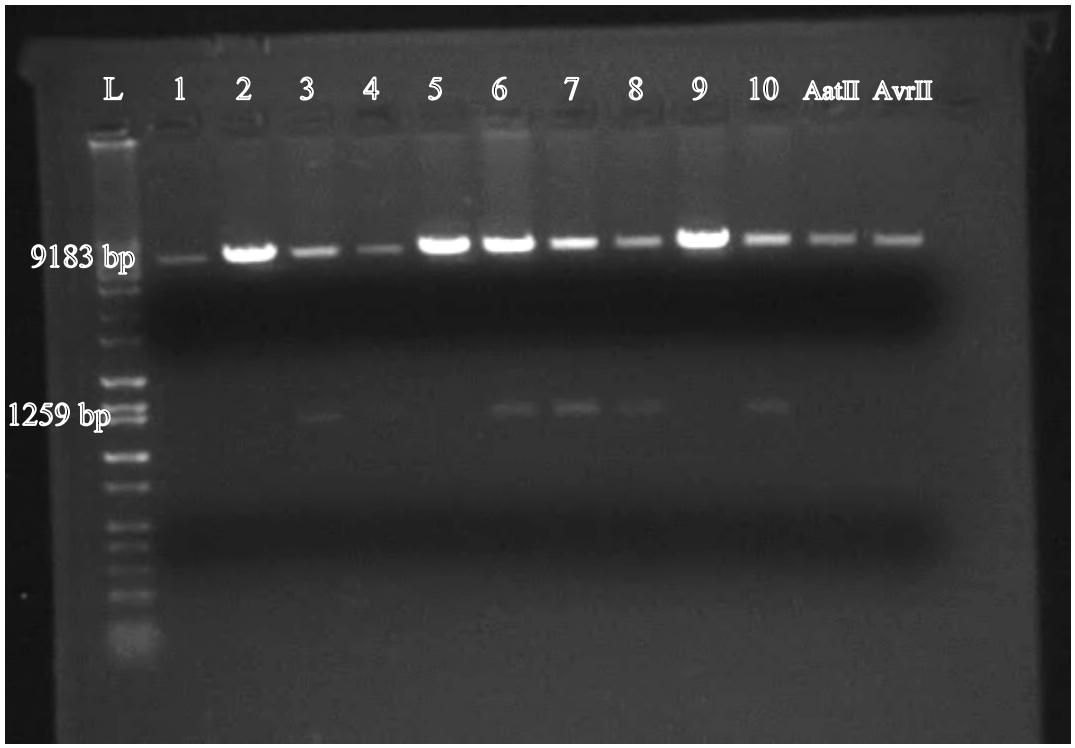


Figure 19. Diagnostic digest of potential clones of pJB027 with AatII and AvrII. L represents ladder, numbers 1-10 represent the 10 colonies chosen from the plates after transformation. Lanes 11 and 12 represent single digests with AatII and AvrII, respectively. Products present at 9138 bp correspond to the plasmid backbone, while product at 1259 bp represent excised insert from successfully cloned eGFP. Gel shows successful clones in lanes 3, 6, 7, 8, and 10.

Figure 19 suggests successful clones of pJB027 in lanes 3, 6, 7, 8, and 10. Select clones were confirmed successful via Sanger sequencing by Eton Biosciences. pJB028, pJB029, pJB030, and pJB031 could all be cloned from pJB027 (see figure 16 for all cloned plasmids). Next, the CAG70 repeats from pCF390 were inserted into pJB027 to create pJB028.

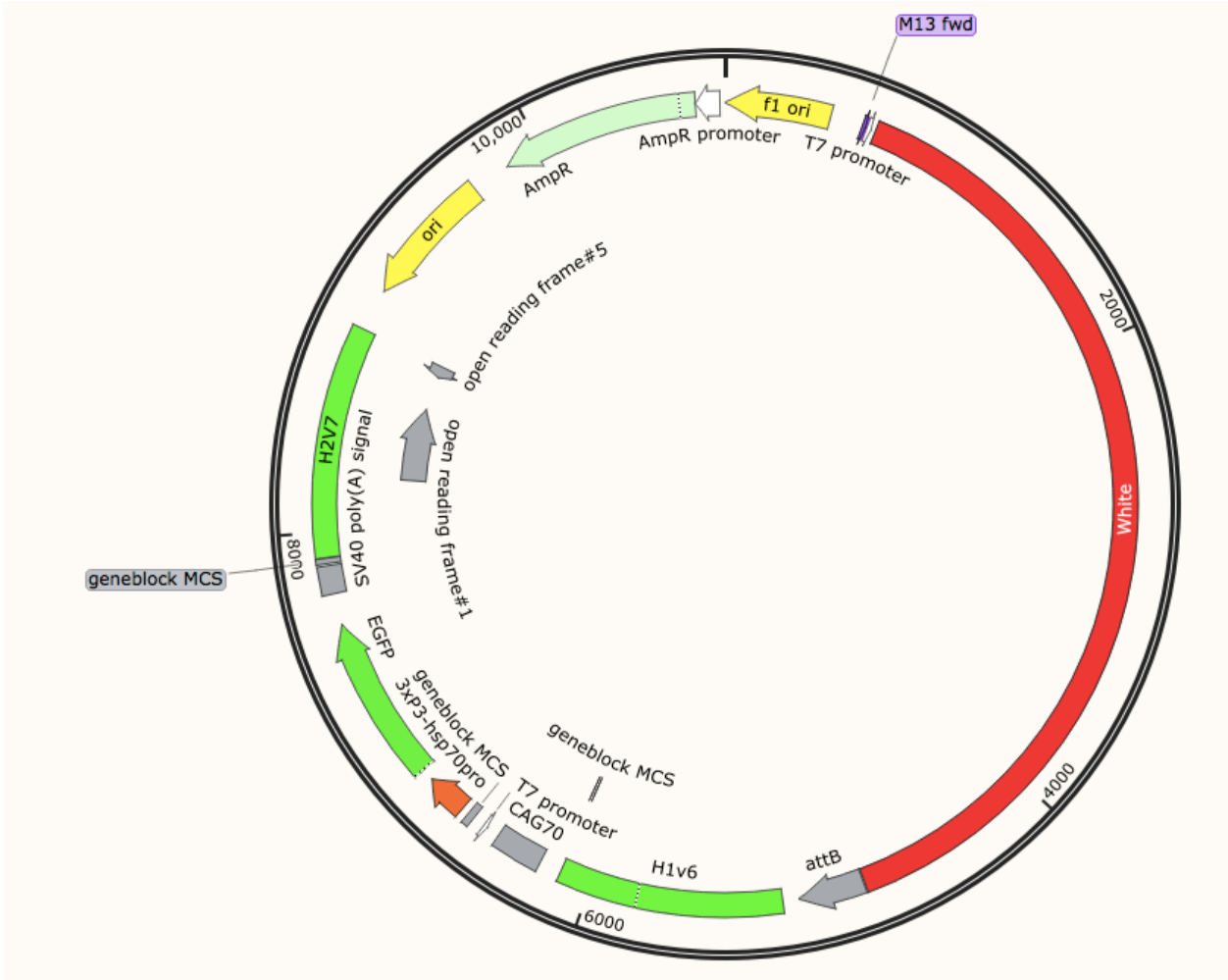


Figure 20. pJB028. pJB027 backbone with CAG70 repeats inserted. Total plasmid size of 10,827 bp.

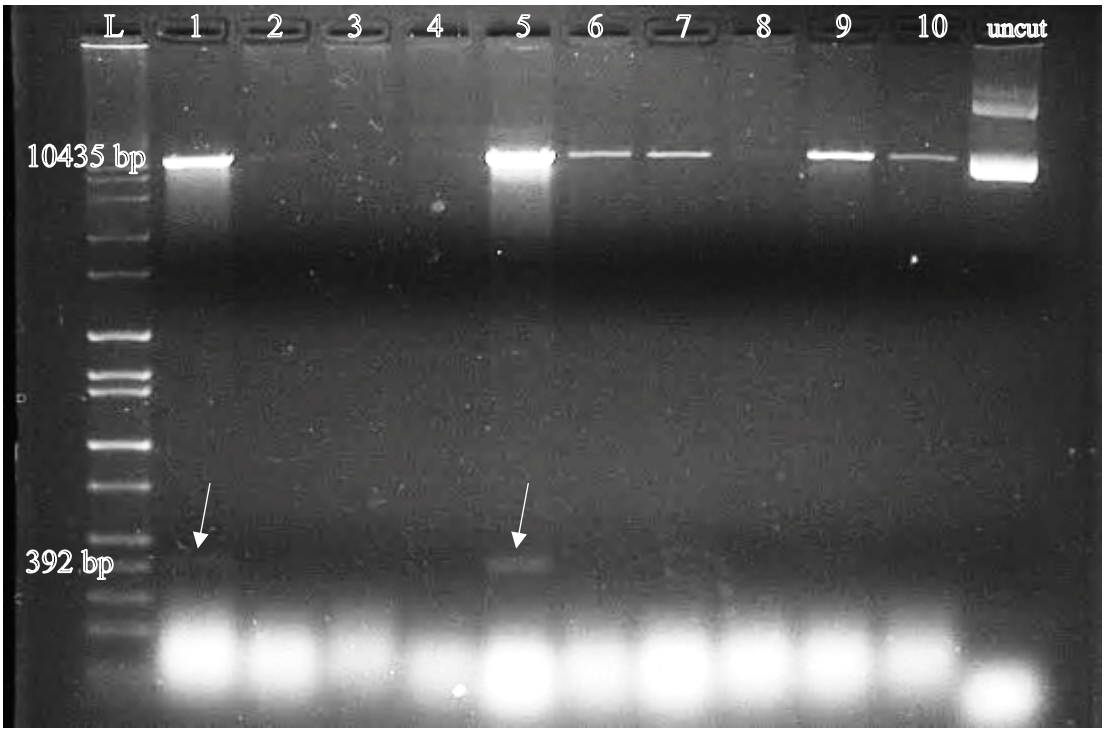


Figure 21. Diagnostic digest of potential clones of pJB028 with PacI. L represents ladder, numbers 1-10 represent the 10 colonies chosen from the plates after transformation. Lane 11 represents the undigested plasmid. Products present at 10435 bp contain the plasmid backbone, while products at 392 bp contain the insert, confirming successful clones in lanes 1 and 5.

The CAG70 repeats were directionally cloned into the pJB027 backbone to create pJB028, as confirmed by sequencing. The backbone of pJB027 was then used to clone CAG130 repeat from pCF391 to create pJB029.

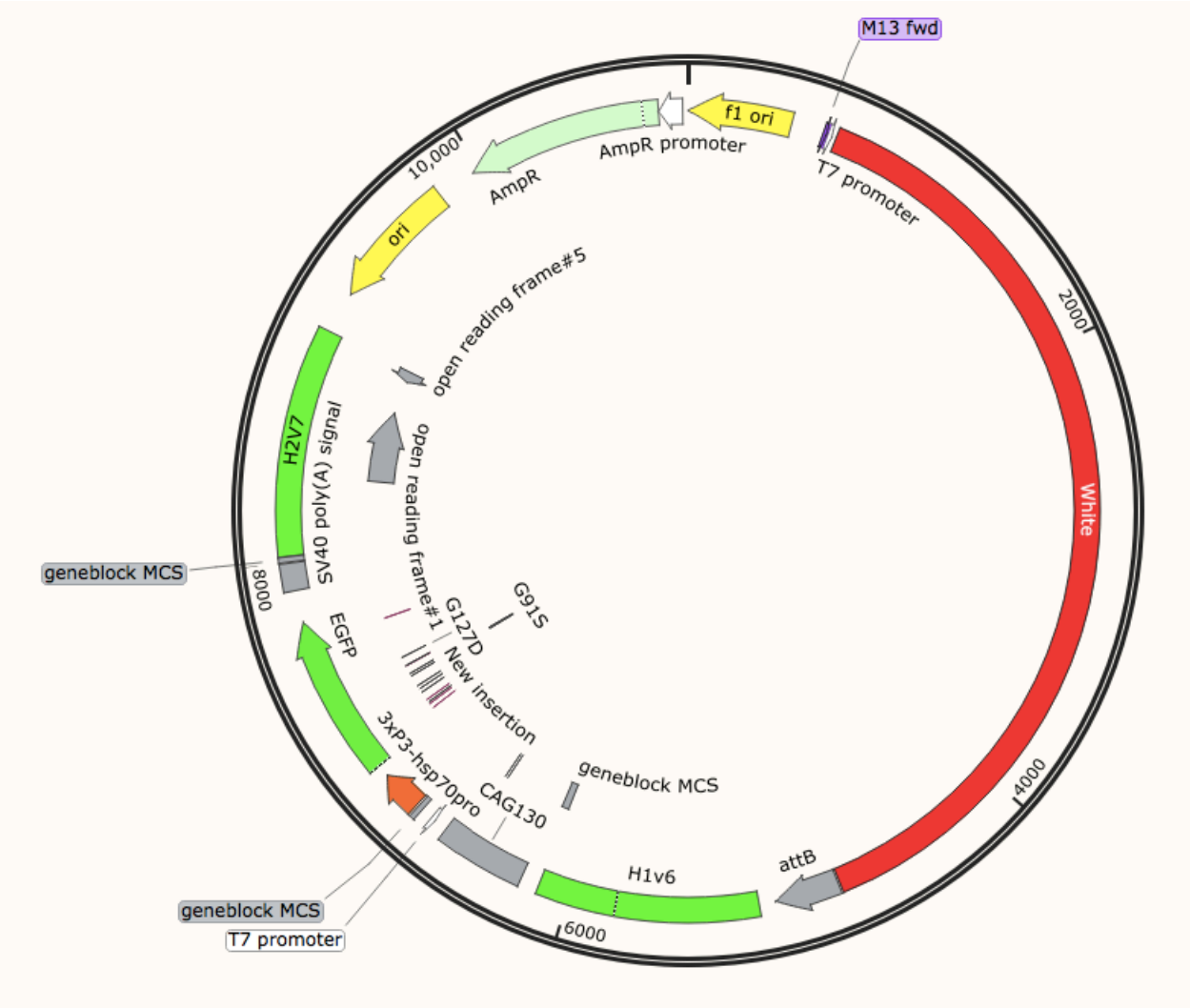


Figure 22. pJB029. Backbone of pJB027 with CAG130 repeats inserted. Total plasmid size of 10,956 bp.

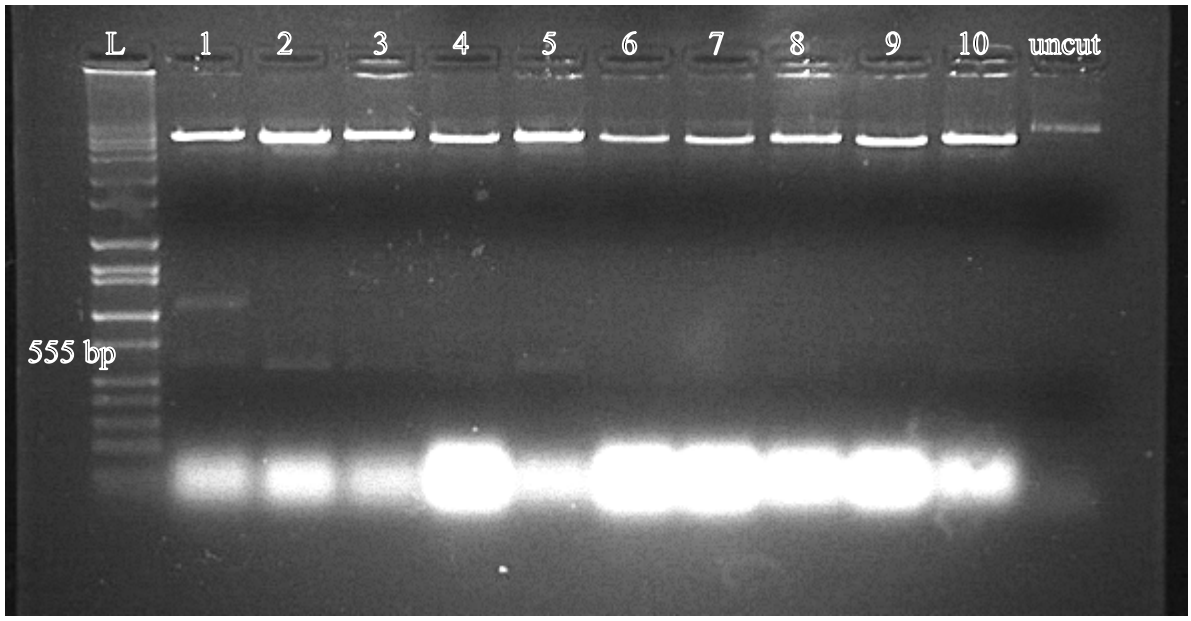


Figure 23. Diagnostic digest of potential clones of pJB029 with PacI. L represents ladder, numbers 1-10 represent the 10 colonies chosen from the plates after transformation. Lane 11 represents the undigested plasmid. Product present at 10435 bp contain the plasmid backbone, while products at 555 bp contain the insert, confirming successful clones in lanes 2-5, 8 and 10. While lane 1 also shows a product at 555 bp, there is a second product running a little over 1 kb, signaling that it should not be used for sequencing, as that was not an expected product for this digest.

The presence of product at 555 bp in the above gel suggests the presence of the repeats, and successful clones of pJB029 were confirmed via sequencing. Lane 1 in the gel also presents with an erroneous product at just above 1 kb in length, and therefore the colony that produced that band was omitted from the pool of possible successful clones.

The next cloning step involved cloning the CAG70 repeats (from pCF390) and the CAG130 repeats (from pCF391) into the pJB026 backbone in order to create pJB32 and pJB33, respectively.

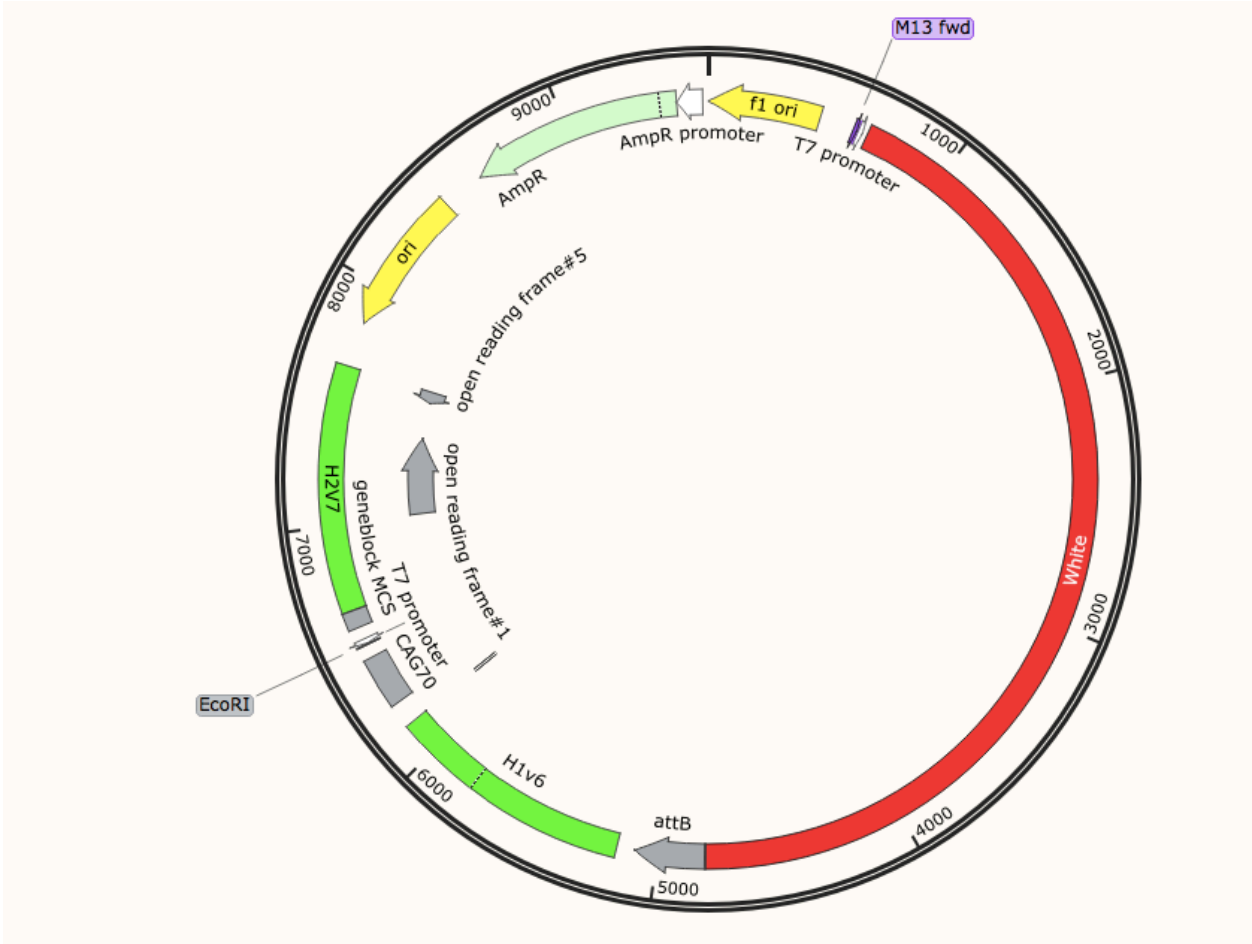


Figure 24. pJB032. pJB026 backbone with insertion of CAG70 repeats. Total plasmid size of 9,586 bp.

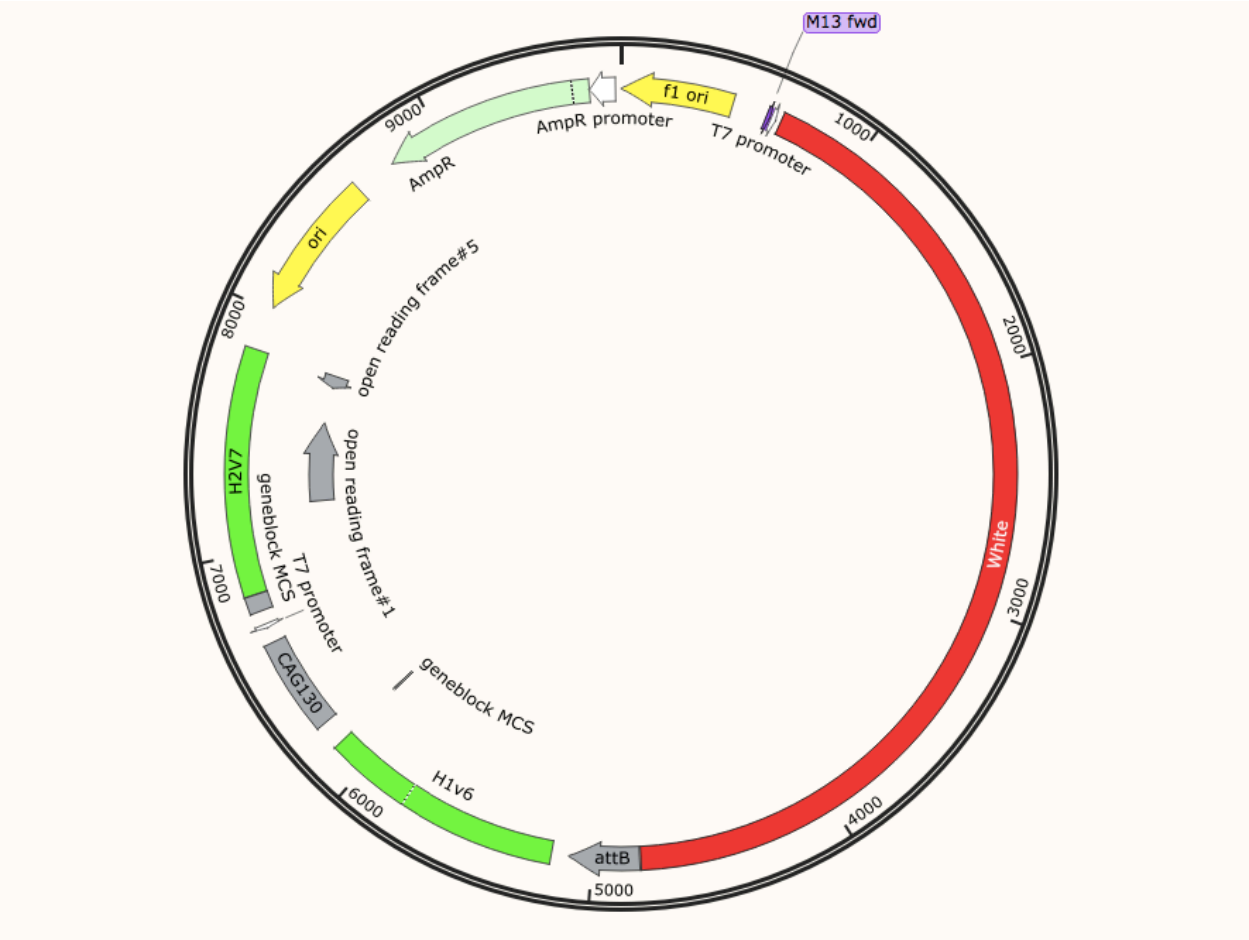


Figure 25. pJB033. Backbone of pJB026 with CAG130 repeats inserted. Total plasmid size of 9,766 bp.

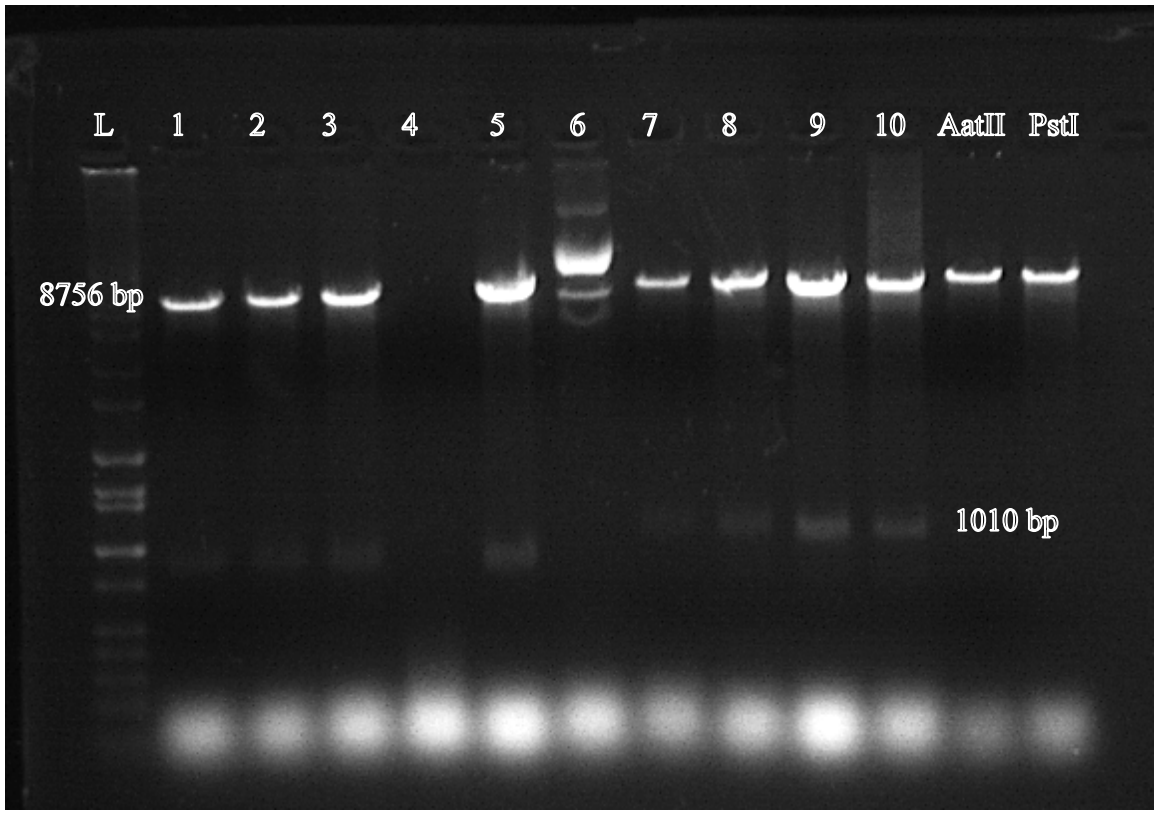


Figure 26. Diagnostic digest of potential clones of pJB032 and pJB033 with AatII and PstI-HF. L represents ladder, numbers 1-5 represent the 5 colonies chosen from the pJB032 plates after transformation, while lanes 6-10 represent the 5 colonies chosen from the pJB033 plates after transformation. Lanes 11 and 12 represents single digests with AatII and PstI-HF, respectively. Products present at 8756 bp contain the backbone plasmid, while products present at 830 bp represent successful clones of pJB032 in lanes 1-3 and 5, while products present at 1010 bp represent successful clones of pJB033 in lanes 7-10.

The lanes with products at 8765 bp for pJB032 and at 1010 bp for pJB033 suggest the presence of the insert, and therefore successful clones. Confirmation of successful clones was determined via sequencing. The CTG70 repeats from pCF590 were then cloned into the pJB026 backbone to create pJB034.

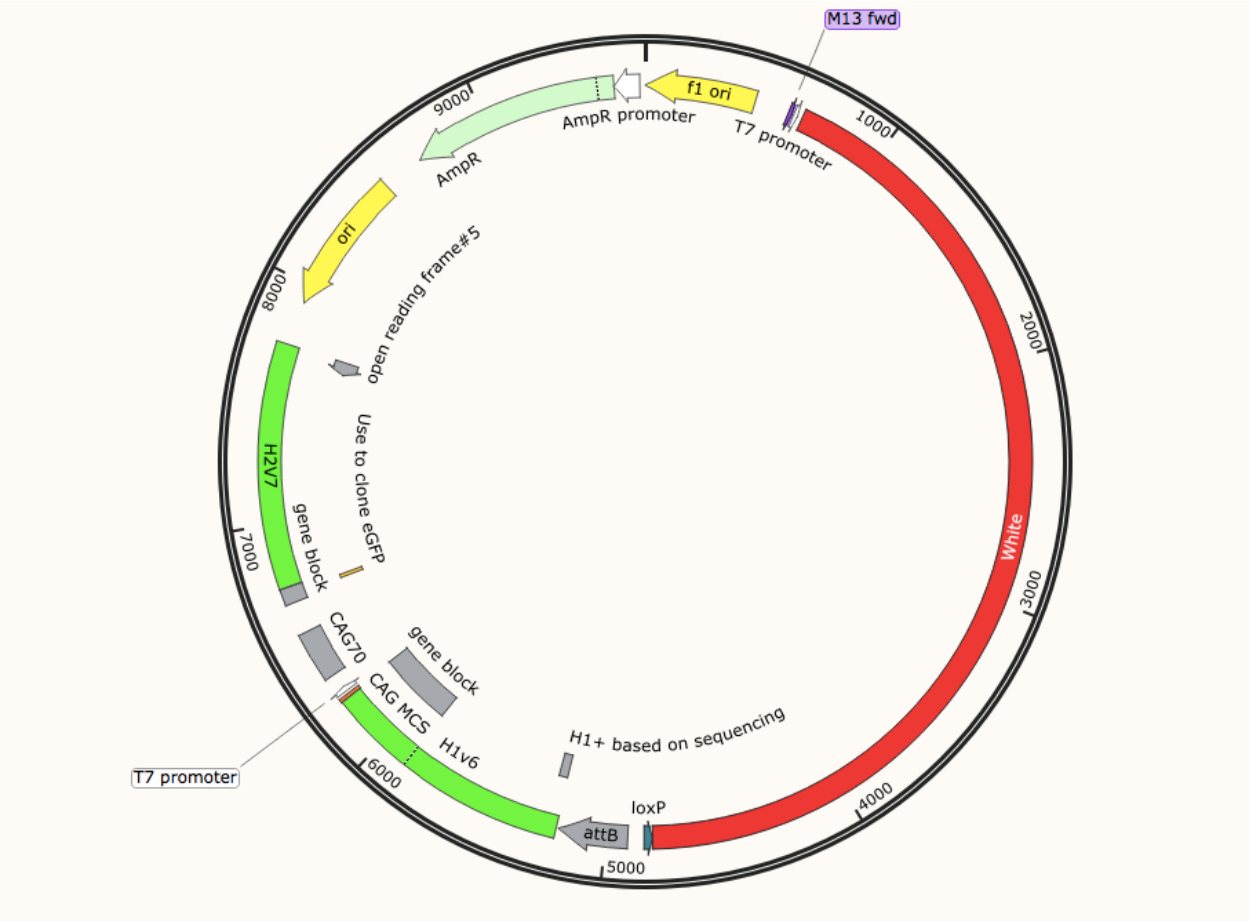


Figure 27. pJB034. Backbone of pJB026 with CTG70 repeats inserted. Total plasmid size of 9,675 bp.

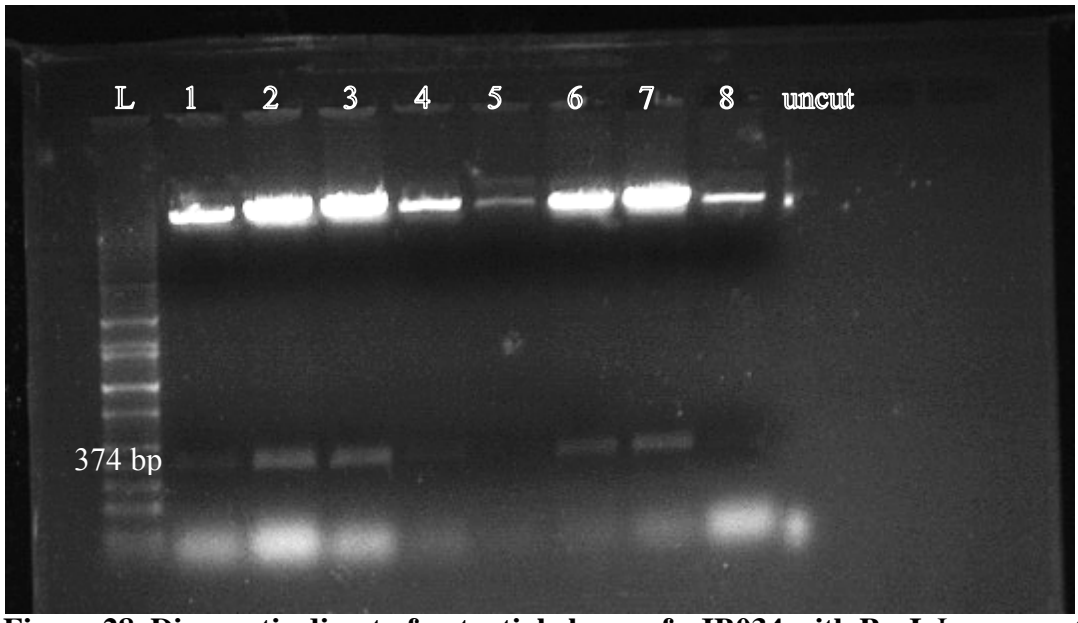


Figure 28. Diagnostic digest of potential clones of pJB034 with PacI. L represents ladder, numbers 1-8 represent the colonies chosen from the plates after transformation. Lane 9 represents the undigested plasmid (not a high enough concentration was added to be visualized on the gel). Products present at 9274 bp contain the backbone plasmid, while products present at 374 bp represent successful clones of pJB034 in lanes 1-4, 6, and 7.

The above gel shows products at 374 bp, suggesting successful insertion of the CTG70 repeats and therefore creation of pJB034, which was confirmed via sequencing. Finally, the CTG70 repeats from pCF590 were cloned into the pJB027 backbone and confirmed via diagnostic digest and gel, followed by sequencing.

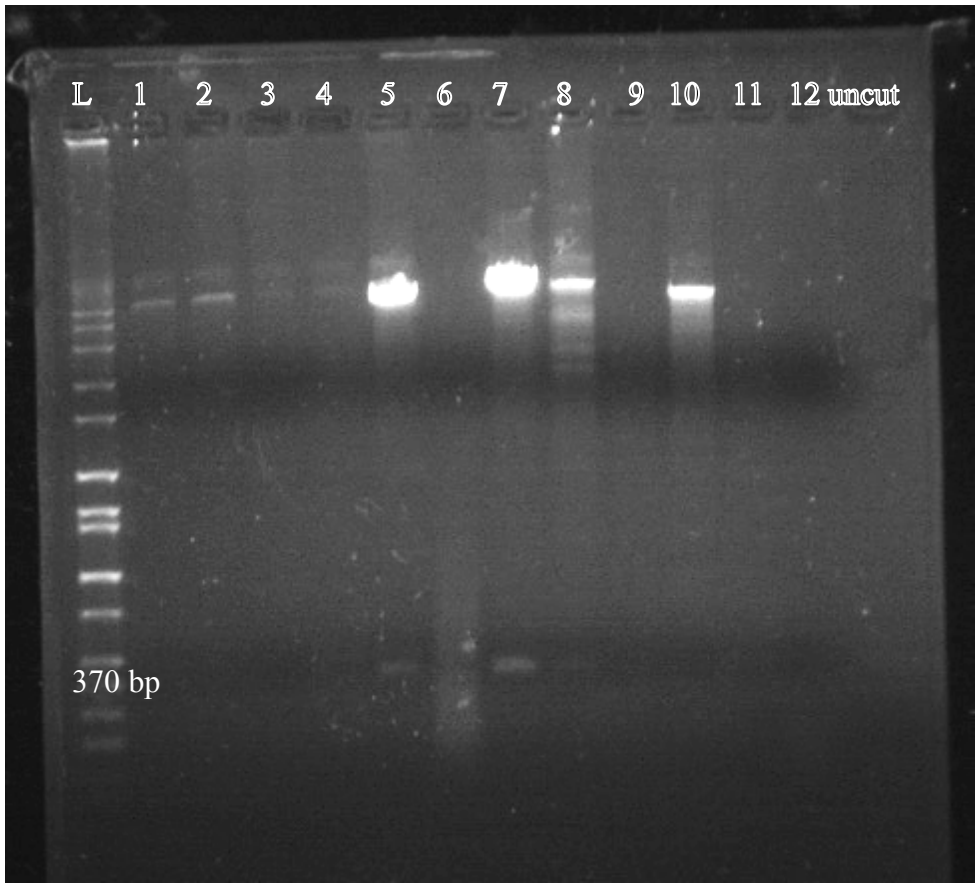


Figure 29. Diagnostic digest of potential clones of pJB030 with PacI. L represents ladder, numbers 1-12 represent the colonies chosen from the plates after transformation. Lane 13 represents the undigested plasmid (not a high enough concentration was added to be visualized on the gel). Products present at 10435 bp contain the backbone plasmid, while products present at 370 bp represent successful clones of pJB030 in lanes 5 and 7.

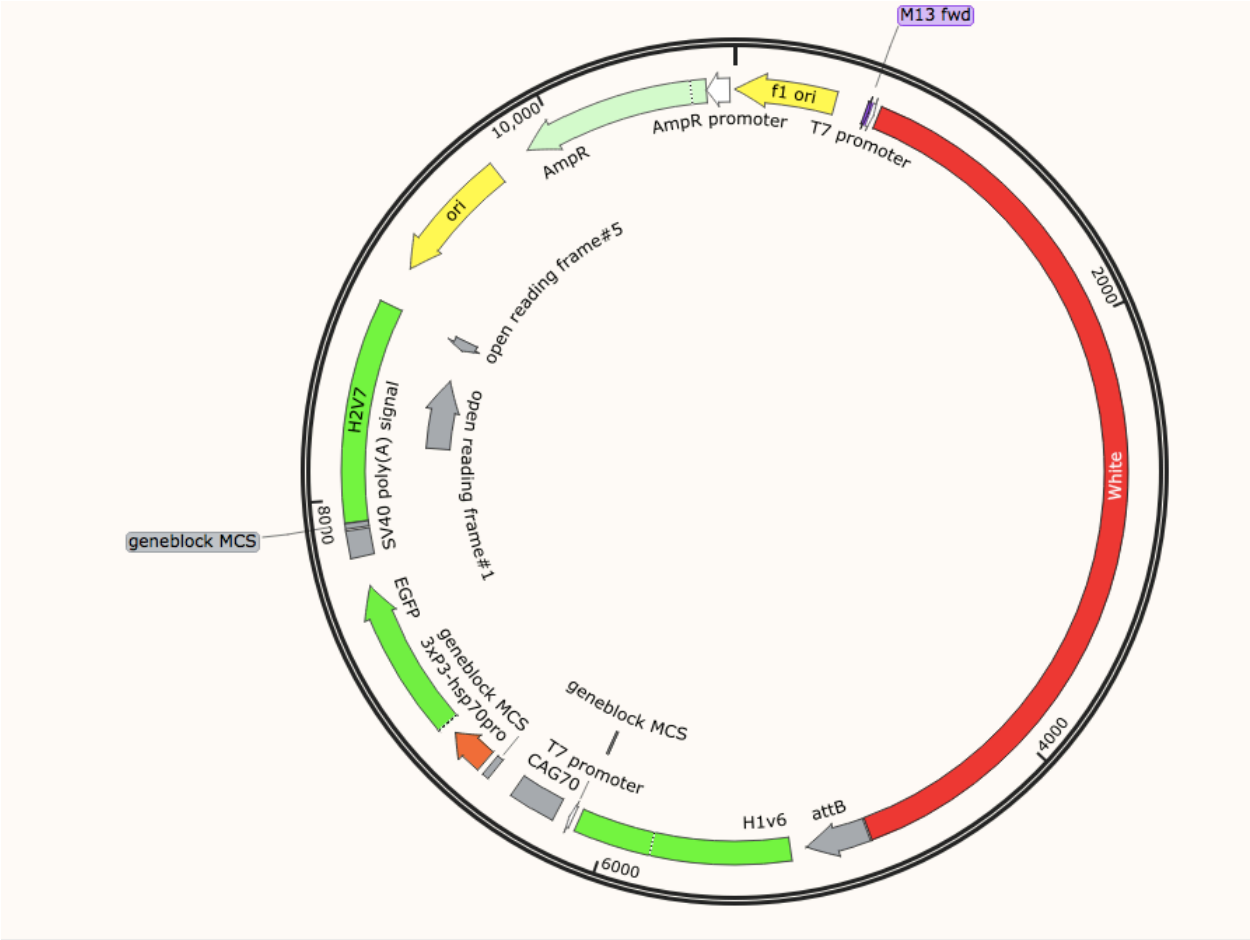


Figure 30. pJB030. pJB027 backbone with CTG70 repeats inserted. Total plasmid size of 10,833 bp.

After cloning, plasmids were created for CAG70, CAG130, and CTG70 inserts into both the pJB26 and pJB27 backbones. The completed plasmids were all saved as glycerol stocks and stored at -80°C. The plasmids containing CAG70 and CTG70 repeats with and without eGFP (pJB0028, pJB0030, pJB032, pJB034) were all injected via BestGene, along with a GFP only (no repeat) control (pJB027). The flies injected with pJB028 (CAG70 with eGFP) were used to characterize the first round of the assay. The CAG130 repeats have been cloned but have not yet been injected, and the CTG130 repeats have not yet been cloned into either pJB026 or pJB027 backbones, as Sanger sequencing of the repeats revealed two templated insertions within the repeats. One of the insertions, located 306 base pairs into the repeats, was already known. The

other insertion, 291 base pairs into the repeats, was not known until the sequencing results were viewed. Because of the new insertion, further cloning with this plasmid (pCF391) was put on hold (see discussion).

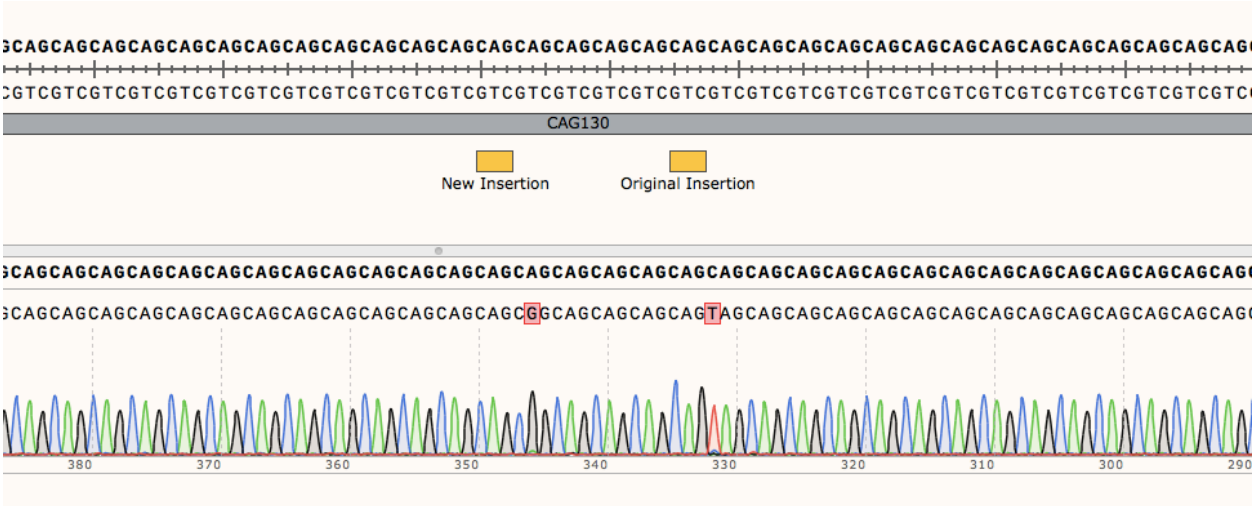


Figure 31. Labeled chromatogram showing two templated insertions within the CAG130 repeats.

PCR of the repeats before injection confirmed that there were actually 71 repeats in the CAG70 and CTG70 plasmids (pCF390 and pCF590, respectively).

After the injected flies from BestGene were obtained, the size of the repeats was once again confirmed. In order to ascertain the number of repeats before they were used as a template for repair in the gap-repair assay, a PCR across the repeats was performed and the products were electrophoresed on a 2% agarose gel at 70 V. M1 and M2 are males containing the CAG 70 repeats before cutting and repair. WT control was a fly prep without the CAG construct, and pCF390 was a plasmid containing the CAG construct. Expected product should run at approximately 300 bp, as is seen in the gel, confirming the size and presence of the repeats.

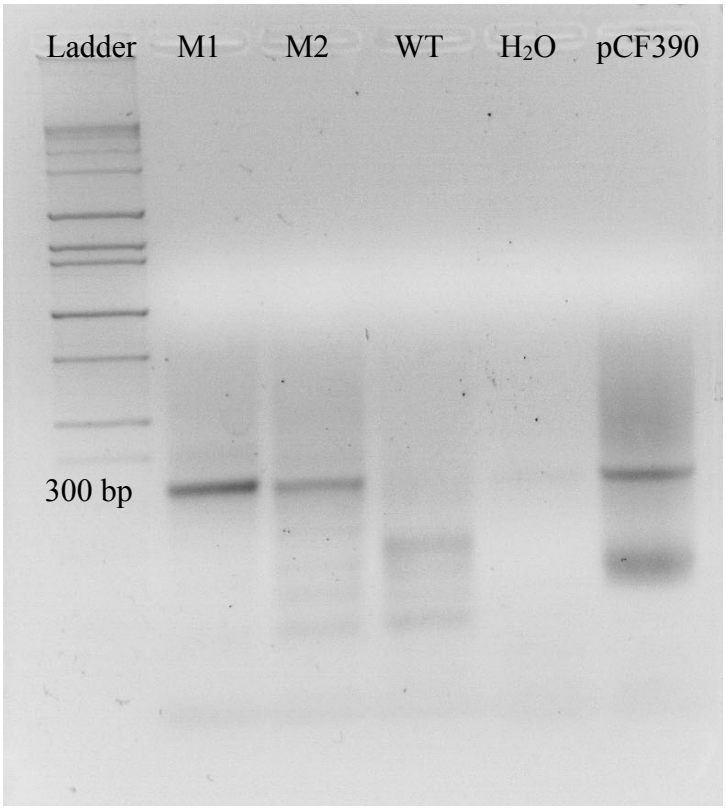


Figure 32. CAG repeat length PCR in first generation males. PCR used primers CAG70FWD and T720B to confirm length of repeats in the first-generation males obtained after injection.

With the confirmation of repeat size *in vivo* completed, the assay could then be performed, and the percent repair events characterized.

Interpretation and Characterization of the Assay

After the induction of the I-SceI-induced DSB, if HR was initiated, invasion was biased to occur from the first homology arm, meaning the elongation would occur across the CAG TNRs first, then continue through eGFP until reaching the second homology arm. Complete HR would therefore result in the loss of DsRed on the recipient, and the gain of eGFP from the donor on the recipient chromosome (see figure 34). It is possible, due to the tract length of synthesis from one homology arm to another and the presence of CAG repeats, that HR could not run to completion and would therefore abort before reaching the second homology arm. Because the

eGFP gene was positioned so close to the second homology arm, any aborted HR would most likely result in no gain of GFP in the repaired recipient chromosome. Therefore, any progeny with fluorescence from GFP was assumed to have completed SDSA. Aborted HR also could result in the loss of DsRed, depending on how far the extension of HR occurred before aborting and end-joining, as a deletion in the DsRed gene would lead to the loss of fluorescence. If faithful canonical end-joining occurred, the *I-SceI* cs would be reconstituted, and as *I-SceI* was constitutively expressed, the *I-SceI* cs would be cut again. If C-NHEJ with mutation occurred, the *I-SceI* cs would then be unrecognizable to the *I-SceI* endonuclease, and the recipient chromosome would retain DsRed. If MMEJ occurred, it is possible that there would be resection into DsRed, resulting in the loss of fluorescence. If there was more minimal resection from MMEJ, the resulting repair event could look identical to repair via C-NHEJ, in that DsRed would remain intact and there would be no gain of GFP. The only repair event in which GFP should be gained on the recipient chromosome would be through HR. Therefore, the type of repair that occurred after the *I-SceI* induced DSB could then be scored by fluorescence. The chromosome on which the repair event occurred, the recipient chromosome, could be selected by the presence of the dominant marker, sternal plural (Sp).

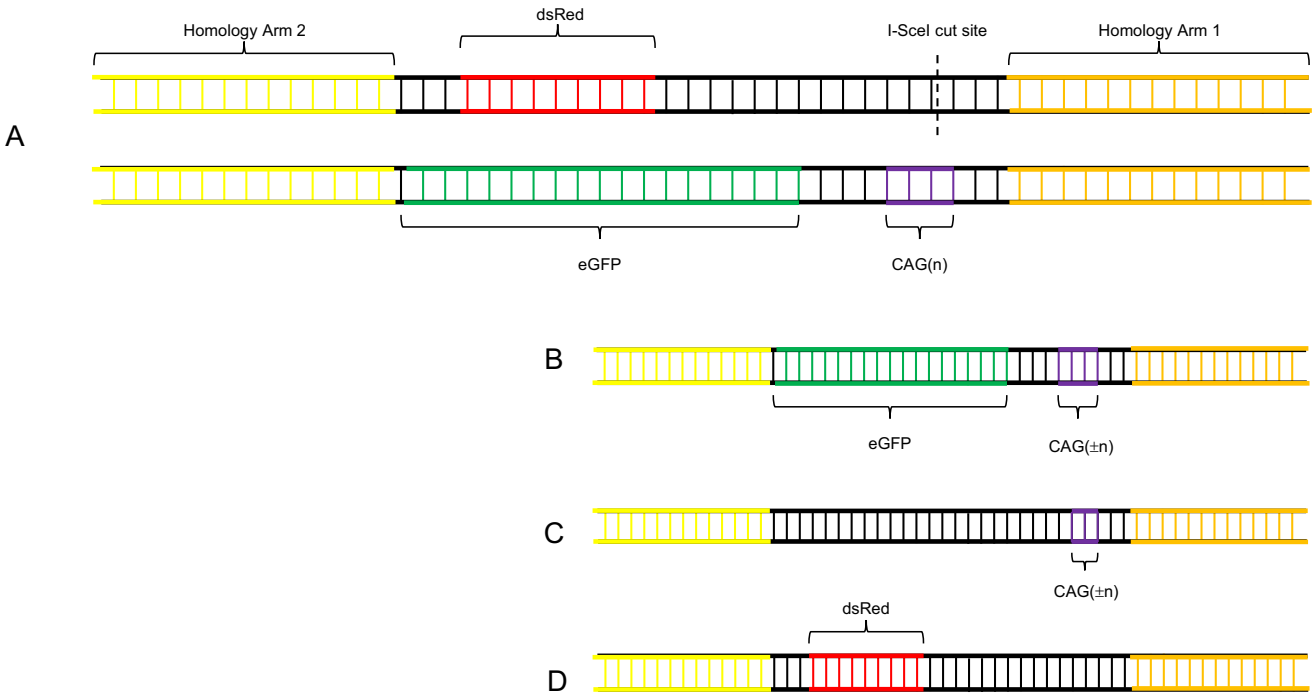


Figure 33. Repair outcomes of the assay. A) Construct of homologous chromosomes in the male germ line in which the *I-SceI*-induced DSB occurs. B-D) Repair outcomes inherited in the progeny. B) Product of repair via complete SDSA. C) Product of repair via incomplete SDSA. D) Product of repair via end joining. Orange and yellow tracts of DNA represent homology arms 1 and 2, respectively. Red represents DsRed, green represents eGFP, and purple represents the CAG TNRs.

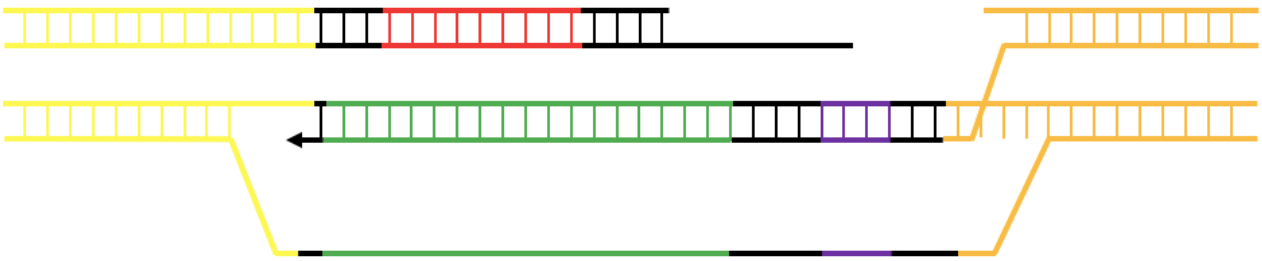


Figure 34. Initiation of repair via SDSA from the first homology arm. Complete SDSA that invades from the first homology arm would first elongate through the repeats, then eGFP, before reaching the second homology arm and reannealing.

All males eclosing from the final cross (see supplemental figure 1) were collected and stored at -20°C. The non-Sp males were collected as a control to study the number of CAG repeats present in the final generation on a chromosome that did not go through repair. All Sp males were also collected to calculate the frequency of each type of repair event (C-NHEJ, MMEJ, complete HR, or aborted HR into EJ). The males that were DsRed-/GFP+ were assumed to have completed HR, as GFP was located only 35 bp away from the second homology arm. These males were further studied in order to characterize the number of CAG repeats present after repair, compared to the starting number before repair. In addition to simply looking at repeat number in the final generation males that repaired via complete HR (DsRed-/GFP+), the flies that were DsRed-/GFP- were also closely studied. The absence of both fluorescent markers indicates that the type repair event that occurred was either MMEJ (with resection far enough into DsRed as to render its protein product non-functional) or aborted HR completing by EJ. The

latter (aborted HR) was of interest in order to help elucidate if the presence of TNRs was affecting polymerase activity during repair, causing HR to abort prematurely.

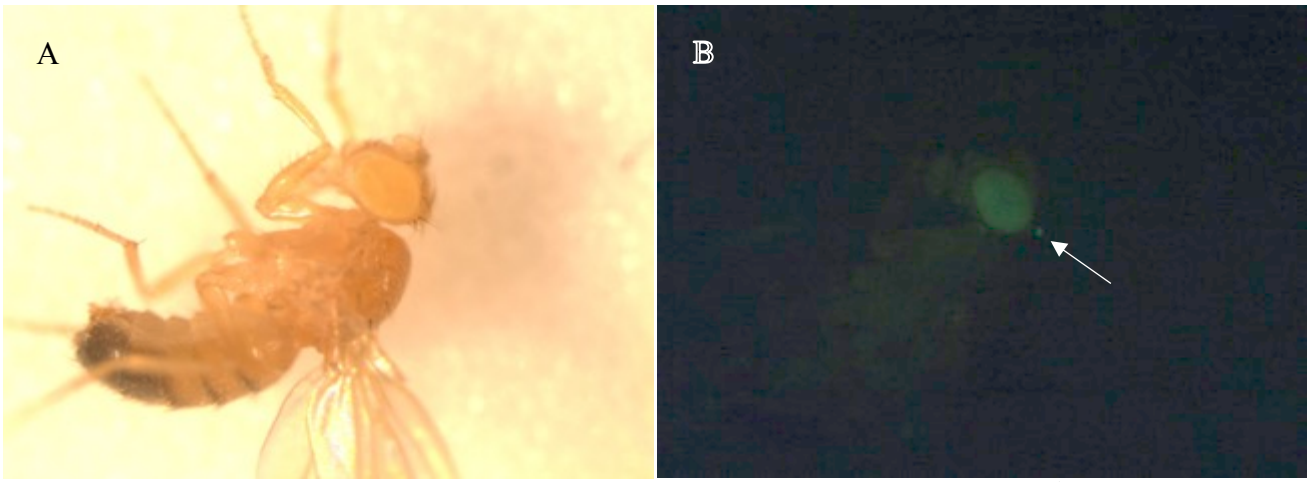


Figure 35. Generation 4 Sp DsRed-/GFP+ male. A. Bright field photo taken at 500 millisecond exposure time at a gain of 1. B. Photo taken under GFP2 filter at gain of 4 and 300 millisecond exposure time. GFP is visible in the eye and ocelli (arrow).

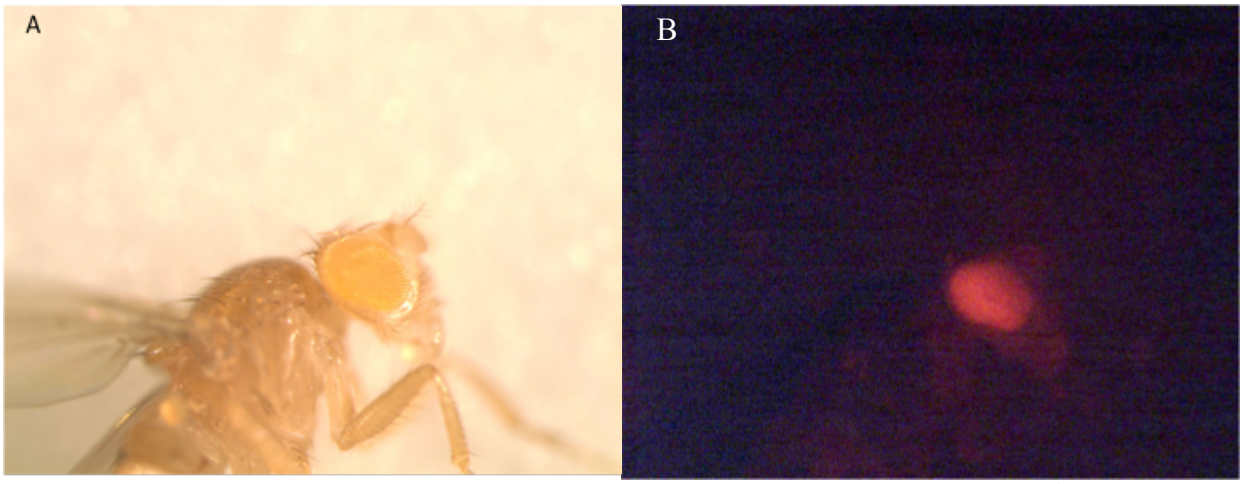


Figure 36. Generation 4 Sp DsRed+/GFP- male. A. Bright field photo taken at 500 millisecond exposure time at a gain of 1. B. Photo taken under DSR filter at gain of 4 and 500 millisecond exposure time. DsRed is visible in the eye.

Repair Outcomes of the Assay

The number of male, Sp progeny from each vial were scored for DsRed-/GFP-, DsRed-/GFP+, and DsRed+/GFP-. These numbers were recorded, and the males collected. Scoring of each vial was done three times over the course of a few days, in order to more accurately reflect any repair events that had occurred, as it was possible that certain repair events could be eclosing earlier or later.

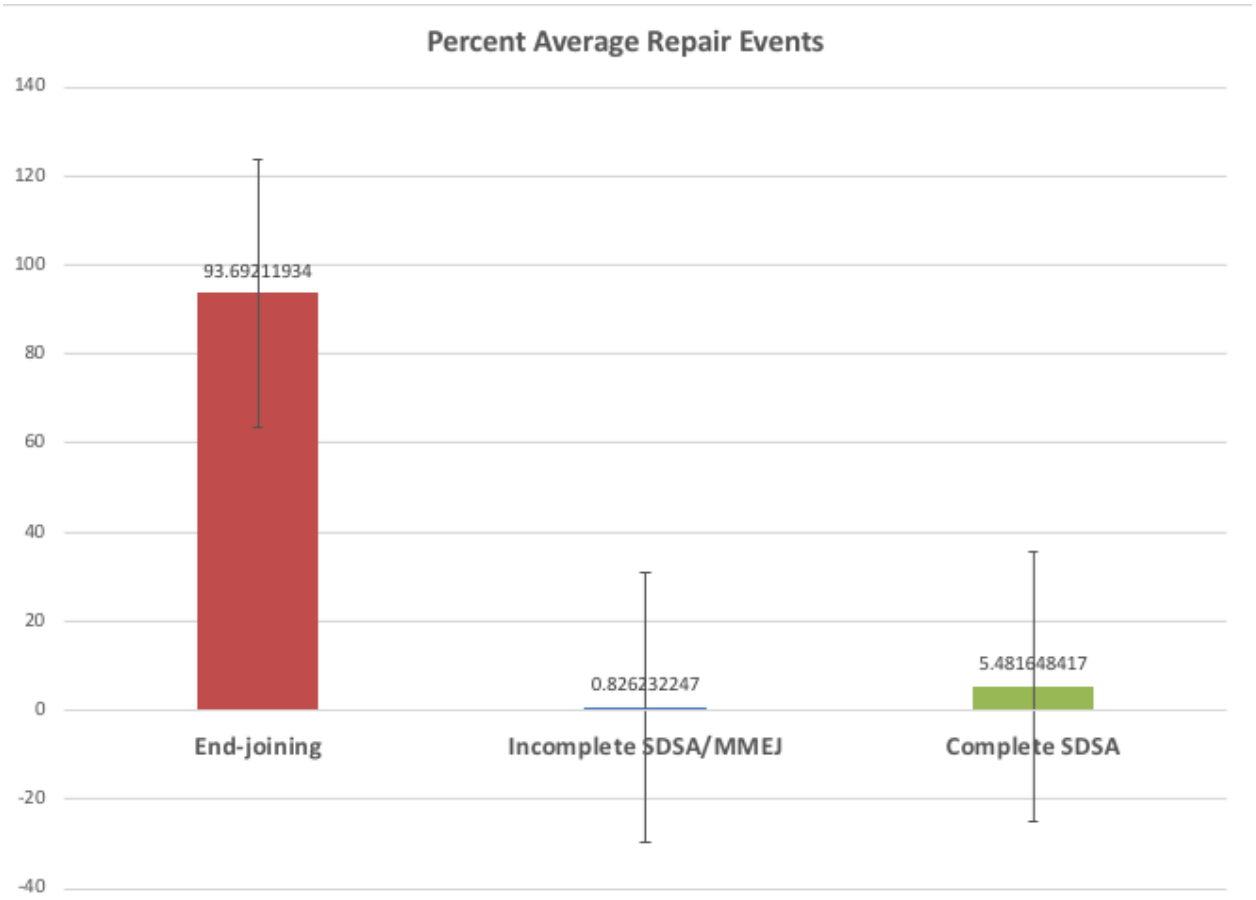


Figure 37. Percent average repair events from the assay with standard error. End-joining represents DsRed+/GFP- progeny, Incomplete SDSA represents DsRed-/GFP- progeny, and Complete SDSA represents DsRed-/GFP+ progeny. 22 vials were used in the scoring.

End-joining encompasses approximately 94% of repair events resulting from the assay. End-joining refers to repair via either canonical end joining or microhomology-mediated end joining (MMEJ) with minimal resection, which would have retained DsRed. Incomplete

SDSA/MMEJ encompasses events of both aborted SDSA resulting in end-joining and MMEJ events with more extensive resection into DsRed, which would have resulted in the loss of DsRed fluorescence. These events comprise less than 1% of all repair resulting from the assay. Complete SDSA refers to repair events that have completed elongation through the homolog and copied all of eGFP and comprises slightly over 5% of the repair events. Each vial was treated as an independent event, and the percentages of each repair event were calculated per vial, then averaged for all vials with standard error.

Discussion

An Inter-Homolog Repair Assay Designed for SDSA Using TNRs as a Template

This assay was specifically designed in order to bias repair via SDSA off of a homolog, in order to get one-ended invasion through the first region of homology (H1). It was important for the assay to preferentially invade on one side (in this case, H1) primarily so we could know which strand of the repeats was being used as a template for repair, as differential orientations of repeats have different instabilities and propensities for expansions and contractions (Freudenreich, 1997). Secondly, in the constructs containing eGFP, it was important for elongation through the repeats to occur before elongation through eGFP. In the event of aborted or incomplete SDSA caused by decreased processivity of the polymerases through the repeats, loss of GFP would occur and we could score that via fluorescence. If the order of eGFP and the TNRs were switched, elongation through eGFP would occur before elongation through the repeats. In the case of aborted SDSA during elongation through the repeats, the repair event would be phenotypically indistinguishable from complete SDSA, as both would have a copy of GFP. The only way to then tell if complete or aborted SDSA occurred would be through PCR, making the screening process much less efficient. Additionally, if mutagenesis were to occur at

an increased rate in the region following synthesis through the repeat, it is possible that GFP could gain a loss-of-function mutation that would be phenotypically observable, but only if GFP were downstream of the repeats. If the GFP gene proceeded the CAG repeats, there would be no such observable phenotype in the event of mutagenesis.

I-SceI was specifically chosen to create the DSB because it was constitutively expressed by a ubiquitin promoter throughout all tissue types and has a large recognition sequence of 18 base pairs. *I-SceI* has also been shown not only to effectively cut in the *Drosophila* genome, but also to lend itself to multiple repair outcomes, including the Rad52-dependent SSA and SDSA pathways (Rong et al., 2003; Preston et al., 2006). The large recognition sequence guarantees that the cut site for this enzyme only occurs once in the genome at the specific locus in the construct where it was engineered. The fact that *I-SceI* is constitutively expressed in all tissues means that presumably, all cells in a single fly expressing both the recipient chromosome with the cut site and *I-SceI* on the X chromosome will have gone through a cutting and repair event. If the break were to repair via c-EJ with no mutation to the cut site, *I-SceI* should continue to cleave the reconstituted recognition sequence until the break repaired via a mechanism in which that sequence was lost. The cutting efficiency can be tested via PCR using primers flanking the *I-SceI* recognition site (see table 5).

Molecular Cloning Resulted in Creation of All but Two Plasmids

There was a delay in cloning with the 130 CAG/CTG repeats (from pCF391) as it was discovered that there were two templated insertions into the repeats. One of these insertions was known, while the other was not. Due to the fact that the nature of how these repeats with the two templated insertions behave has not yet been characterized, we went back to the original CTG 130 plasmid, pGEM$(CTG)_{130}$ (as characterized in Freudenreich et al., 1997). Because the MCS

was designed with the original CAG 130 plasmid backbone in mind, the pGEM backbone was lacking in compatible enzymes for cloning. In order to successfully clone the CTG 130 repeats into the pJB026 and pJB027 backbones, both sides of the repeat will need to be digested with blunt-cutting enzymes, resulting in non-directional cloning of the insert. Both pJB backbones could be digested with PmeI and PshAI, while the repeats would be digested with StuI and PvuII. Digesting the plasmids with blunt-cutting enzymes will decrease the ligation efficiency as well as cause the insert to be non-directionally cloned. The plasmids cloned with CAG130 inserts (pJB033 and pJB029) also contain the two templated insertions, and therefore were never sent out for injection, but were still stored at -80 °C.

The Vast Majority of Repair is Comprised of End Joining Events

According to preliminary data, approximately 94% of repair events appear to be either canonical end joining or microhomology-mediated end joining that did not resect into DsRed, as evidenced by the percentage of Sp, male progeny that were DsRed+ in the final generation (see figure 37). From the paper in which the *I-SceI* position and promoter used in our assay was characterized, nearly 100% of progeny exhibited cutting by *I-SceI* (Preston et al., 2006). It is therefore likely that the majority of the DsRed+/GFP- events shown in this assay are indeed the product of some type of end joining event, and not that of an intact cut site, which can be tested and confirmed via PCR. Additionally, this same paper indicates that the two repair pathways (HR and NHEJ) compete with one another in DNA repair. When there is an acceptable template on the homolog, gene conversion competes with EJ, comprising approximately 19.4% of the repair events. This percentage is higher than what was observed in our assay, indicating that the movement of our *I-SceI* cs closer to the region of homology may prove to be very beneficial (as discussed in the following section). Additionally, Preston et al. observed that the majority of

repair events were SSA (61.3%). Phenotypically, a repair event via SSA would be indistinguishable from MMEJ with deletions into DsRed and aborted SDSA. A similar result is not observed in our assay, where the smallest percentage of repair events presented as DsRed-/GFP- (~1.13%). Perhaps there are fewer sequences of longer homologies located on the recipient chromosome that biases repair away from SSA and MMEJ, or there is another difference between the two assays preventing this type of repair.

Future Iterations of the Assay will be Further Biased to Repair Via SDSA

Because the most interesting data that can be gathered from this assay are the result of either complete or aborted SDSA, it would make successive rounds of the assay much more efficient if we could increase the rate of strand invasion, leading to increased rates of both complete and incomplete SDSA. The more the assay is biased to repair via HR, the more useful results can be obtained from the assay. This assay relies on inter-homolog repair instead of repairing off of a sister chromatid. While two sister chromatids should be identical, and therefore this repair choice is genetically silent and higher fidelity than inter-homolog repair (which can often lead to loss of heterozygosity) this appears to not be the reason for preference of repair via sister over the homolog in many species (Kadyk et al., 1992). Sister chromatids are physically closer to one another during S and G2 phases, which could increase the ratio of inter-sister repair as compared to inter-homolog repair (Kadyk et al., 1992; Moynahan et al., 2010). However, in *Drosophila*, the homologous chromosomes are mitotically paired (Steven et al. 1908), and therefore the incidence of inter-homolog repair is more common, although still quite rare (Rong et al., 2003). This study by Rong et al. in 2003 observes approximately 2% DSB repair via HR using the homolog as a template when there is no homologous sequence in the vicinity of the DSB. This same study also asserts that HR off of a homolog seems to be much more favored if

the region of homology is directly flanking the cut site. They observed that when this is the case, ~65% of repair events are from inter-homolog repair, outcompeting both HR using the sister as a template and end joining. Additionally, because *I-SceI* is such an efficient cutter, repair off of the sister may be disfavored, as both sister chromatids could contain a DSB. The efficiency of *I-SceI* cutting could also bias our assay away from repair off of the sister because that type of repair would reconstitute the *I-SceI* cs, leading the site to get cut once again, until it repairs by a mechanism that leads to the loss of the cut site.

The first step we are taking to further bias the assay to repair via HR is to reengineer the recipient chromosome in order to move the I-SceI cut site closer to H1. Currently, the I-SceI cs is 285 bp away from H1. The closer the I-SceI cs is moved towards H1, the less resection has to occur before reaching the region of homology, therefore increasing the odds that the break will repair via HR. Additionally, it may be possible to find a CRISPR target site near H1 and create a double-strand break via Cas9. A stock of flies expressing Vasa-Cas9 (a germline-specific promoter of Cas9) could be used for the assay in order to drive repair events in the germline that would be inherited in the next generation. However, using CRISPR-Cas9 in *Drosophila* to create DSBs could lead to off-target effects where the gRNAs lead the Cas9 endonuclease to cut at loci other than the intended target (Zhang et al., 2015). While it is possible that using CRISPR could be faster than redesigning the recipient homolog to have *I-SceI* closer to H1, CRISPR would be a less specific tool for creating a DSB than using *I-SceI*. Additionally, DNA repair in *Drosophila* based off of an *I-SceI* cut has been well characterized and has a track record of success.

Aside from moving or changing the cut site, there are several mutant backgrounds that could increase the rate of HR by decreasing the rates of end joining. In order to determine the best way to increase the rate of strand invasion in a mutant background, it must first be

determined whether canonical end-joining or MMEJ without deletion of DsRed predominated the recovered DsRed+ events. The predominant mechanism of repair can be determined via PCR and Sanger sequencing (see table 4). Because *I-SceI* is ubiquitously expressed, it would be surprising, although still possible, to see an event in which the cut-site was still in-tact. Therefore, if the break repaired via c-EJ, it is likely that the repair was mutagenic, and the cut-site was either lost or somehow changed so that *I-SceI* could no longer recognize the sequence. Because there is little DNA lost in the break during c-EJ, primers closely flanking the cut site can be used (See figure 11), and the presence of a product would confirm repair via c-EJ. Alternatively, repair via MMEJ involves more extensive resection, and therefore primers would produce differently sized products depending on both the type of repair, and, if the repair was via MMEJ, the amount of resection. A larger product would imply less resection.

If PCR indicated that the majority of the DsRed progeny were repaired via c-EJ and not MMEJ, as is expected since c-NHEJ is more common than Alt-EJ in *Drosophila* (Johnson-Schlitz et al., 2007), then the most likely next step would be to repeat the assay in a ligase IV-deficient background. Ligase IV is necessary for repair via canonical end-joining, as it ligates the two ends of the break back together (Grawunder et al., 1997). Therefore, if the generation of *Drosophila* completing the repair events were mutants for ligase-IV, they would be unable to repair double-strand breaks via c-EJ. This would result in either the other repair pathways compensating for this loss of one type of repair, such as HR, SSA, or MMEJ (all of which require resection) or in an increase in DNA damage that goes unrepaired. Unrepaired DNA DSBs would lead to loss of a portion of the chromosome after the next round of cell division, meaning any unrepaired DSBs in the germline would likely result in non-viable offspring (Rong et al., 2003). We would therefore be able to observe that the other repair pathways were not

compensating, as we would obtain fewer progeny. This, however, does not seem likely, as it has been shown that when one repair pathway is defective in *Drosophila*, the other repair pathways compensate (Johnson-Schlitz et al., 2007; Preston et al., 2006). While the incidence of homologous recombination would likely increase, it is possible that MMEJ could also take over as the dominant form of repair. Alternatively, if PCR analysis proves that MMEJ dominates over canonical end-joining, then repeating the assay in a polymerase theta deficient background could decrease the frequency of MMEJ. Polymerase theta has been shown to promote MMEJ by annealing microhomologies of a DSB after resection (Beagan et al., 2017). Therefore, the pol theta mutant would do nothing to inhibit repair via canonical EJ, but if the break started to resect, it would be unable to repair via MMEJ, and would have to repair via SSA or HR. Polymerase theta mutant flies are viable, but not incredibly healthy, as evidenced from their 4.97% hatching frequency (as compared to 85.1% in the wild type control) and therefore a mutant of ligase IV would be more feasible for this assay than a polymerase theta mutant (Alexander et al., 2016). Additionally, a ligase IV/polymerase theta double mutant would also not a plausible mutant to study in subsequent iterations of the assay, as that same study found 0% hatching frequency in a lig4/pol theta double mutant. Either of these two mutant backgrounds could potentially dramatically increase the number of SDSA repair events, and therefore also increase the amount of data collected on synthesis tract lengths and trinucleotide repeat instability after Rad52-dependent repair.

Future Directions and Conclusions

Aside from generating mutant backgrounds in order to increase the rate of repair via SDSA, future experiments will include characterizing repeat instability in several different mutant backgrounds. As evidenced from previous work in the McVey lab on another DNA DSB

repair assay (the P{w^a} assay, see McVey, 2003), in mutant backgrounds of both of the translesion synthesis (TLS) polymerases eta and zeta, there is a significant decrease in repair via HR (45% and 50%, respectively) as compared to the wild type (Kane et al., 2012). The overall decrease in HR events, but no significant change in synthesis tract lengths from these mutants suggest that polymerases eta and zeta may play a role in the initiation of HR, but that elongation is likely due to the normal replicative polymerase delta. Mutations to the normal replicative polymerases, delta and epsilon, have been shown to significantly increase TNR expansion rate in yeast as compared to the wild type, suggesting that these polymerases play a role in stabilizing repeats (Shah et al., 2012). It has been shown that when faced with fragile sites in the genome (such as TNRs) that stall the normal replicative polymerases, there can be a polymerase switch to more specialized polymerases, such as eta or kappa (Barnes et al., 2017). While polymerase kappa is not present in *Drosophila*, studying the effects of a mutated TLS polymerase, such as eta or zeta, could provide interesting results as to the role of polymerase switching and the use of TLA polymerases to replicate through TNRs during repair.

In addition to the assay using GFP and CAG70, the assay will be repeated with different numbers of repeats, different orientations of the repeats, and in the absence of GFP. All constructs have a control without GFP present in the donor chromosome (plasmids created from the pJB026 backbone). While the absence of GFP will make scoring the assay much more difficult, it is an important control in reducing the length of synthesis needed for complete SDSA. While aborted SDSA has so far only constituted a very small minority of events (less than 1%), implying that when the cell begins repair via SDSA, it more often than not completes it, it is possible that in future iterations of the assay with longer repeat lengths, there will be more incomplete SDSA observed. The total amount of synthesis from the first to the second homology

arm when GFP is included is slightly under 2 kb in length, while GFP by itself is 1218 bp. By removing GFP from the donor construct, the total length of synthesis goes down to approximately 800 bp. P{w^a} data shows that in a wild type background, 80% of repair events via SDSA complete synthesis tract lengths of at least 0.9 kb in length and 69% show synthesis tract lengths of at least 2.4 kb in length (McVey, 2010). Therefore, synthesis across two kilobases of synthesis is not an improbable length, but having a GFP- control is still important, as a similar rate of complete SDSA would point to the length of repair not being an issue. Aborted SDSA then could be due to a number of reasons, one of which being the presence of the trinucleotide repeats.

We also have obtained a stock of flies injected with a GFP only control (lacking any CAG repeats on the donor chromosome) to control for any effects of the TNRs. The results of the assay with these flies could indicate any effects the CAG/CTG repeats have on repair pathway choice. If there is a lower rate of incomplete SDSA in the GFP only constructs, it would imply that the TNRs are causing SDSA to abort before completion of elongation. Finally, there are control flies that have both the donor and recipient chromosome, but do not express *I-SceI* in the germ line, and therefore we are able to observe the innate instability of the trinucleotide repeats without the induction of a double-strand break. These results will give us a background level of their instability *in vivo*.

While the assay, as it stands, has been done using a CAG70 construct, this is a fairly stable number of repeats, as a polyglutamine tract of 75 CAG repeats has been shown to be stable in somatic as well as germline tissue in *Drosophila* (Jackson et al., 1998). Using a longer repeat length of 130 CAG repeats, and possibly even more, could demonstrate the repeat-length dependence on TNR instability. With more repeats, it has been shown that larger expansions can

occur, thereby increasing the number of repeats at a faster rate (Petruska et al., 1996). The longer repeats are expected to show a greater baseline instability without the induction of a DSB, but also are expected to cause more expansions during repair than the constructs with only 70 repeats. The longer length of the repeats could also cause more incomplete SDSA, as hairpin formation and other secondary structures on the template strand during elongation could cause dissociation of the polymerase, leading to aborted SDSA and resulting in end-joining. This result would be interesting in and of itself in that if longer tracts of repeats caused more aborted HR and end-joining, there could also be a larger number of mutagenic events occurring in the vicinity of the repeats.

Repeats will also be inserted into the construct in the reverse orientation, so that the template strand for SDSA will be CTG repeats instead of CAG. CTG repeats form more stable hairpins (Petruska et al., 1996; Hartenstine et al., 2000), and therefore when present on the template strand, could lead to more contractions of the repeats instead of expansions. The results of the assay from CAG and CTG repeats of corresponding size could be used to directly compare which orientation of the repeats is more stable in the genome. In yeast, it has been shown that the orientation of CAG/CTG repeats is directly correlated with instability. When the repeat orientation places CAG on the lagging strand, the repeats are more stable, showing few changes to repeat number after replication, with the exception of a couple expansions. The CTG repeats on the lagging were less stable overall, showing a significant frequency of contractions (Freudenreich et al., 1997). While there is no leading or lagging strand during DSB repair, it could be hypothesized that when the strand serving as the template for repair contains CTG repeats, more secondary structures would form, leading to contractions. Conversely, when CAG

repeats were on the template strand, the newly synthesized strand would contain CTG repeats, making it more likely to form hairpins, thus leading to expansions.

From this study, we were able to design a gap-repair assay that was able to repair an *I-SceI*-induced double-strand break off of the homolog. From this assay, we could determine the rates of each repair event, showing that end joining was the dominant repair pathway for this construct. Future experiments will determine the stability of trinucleotide repeats across generations of *Drosophila*, as well as the stability of the repeats after repair via SDSA. These experiments will be performed in a variety of mutant backgrounds and with several different constructs, as discussed above, in order to further elucidate the molecular mechanisms underlying TNR instability as it is related to DNA DSB repair in a metazoan system.

Acknowledgments

I would like to thank Dr. Mitch McVey for mentoring me, always taking the time to answer my questions, and for letting me conduct research in his lab. I am so thankful to have had this research opportunity during my time as an undergraduate; it has helped me gain insight into my future career aspirations and has allowed me to further pursue my passion for science. The McVey Lab has been the highlight of my undergraduate career. I would also like to thank Dr. Sarah Dystra for her continued support and mentorship throughout this project, as well as for her role in designing and executing the assay. Additionally, I would like to thank Dr. Catherine Freudenreich for her roles as both my advisor and thesis committee member. Lastly, thank you to Dr. Alice Miller, Keya Viswanathan, and the rest of the McVey Lab family for their constant support and for always making lab feel like home.

References

1. Alexander J, Beagan K, Orr-Weaver T, McVey M (2016) Multiple mechanisms contribute to double-strand break repair at rereplication forks in *Drosophila* follicle cells. Proc Natl Acad Sci USA. 113(48): 13809-13814.
2. Barnes R, Hile S, Lee M, Eckert K (2017) DNA polymerases eta and kappa exchange with the polymerase delta holoenzyme to complete common fragile site synthesis. DNA Repair. 57: 1-11.
3. Beagan K, Armstrong RL, Witsell A, Roy U, Renedo N, Baker AE, Shärer OD, McVey M (2017) Drosophila DNA polymerase theta utilizes both helicase-like and polymerase domains during microhomology-mediated end joining and interstrand crosslink repair. PLoS Genet. 13(5): e1006813.
4. Budworth H, McMurray C (2013) A Brief History of Triplet Repeat Diseases. Methods Mol Biol. 1010: 3-17.
5. Freudenreich C, Stavenhagen J, Zakian V (1997) Stability of a CTG/CAG Trinucleotide Repeat in Yeast Is Dependent on Its Orientation in the Genome. Molecular and Cellular Biology. 17(4): 2090-2098.
6. Fu J, Kanno T, Liang S, Matzke A, Matzke M (2015) GFP Loss-of-Function Mutations in *Arabidopsis thaliana*. G3: Genes, Genomes, Genetics 5(9): 1849-1855.
7. Gonzales E, Yin J (2010) Drosophila Models of Huntington's Disease Exhibit Sleep Abnormalities. PLOS Currents Huntington Disease. Edition 1 doi: 10.1371/currents.RRN1185.

8. Grawunder U, Wilm M, Wu X, Kulesza P, Wilson T, Mann M, Lieber M (1997) Activity of DNA ligase IV stimulated by complex formation with XRCC4 protein in mammalian cells. Nature 388: 492-495.
9. Hartenstine M, Goodman M, Petruska J (2000) Base Stacking and Even/Odd Behavior of Hairpin Loops in DNA Triplet Repeat Slippage and Expansion with DNA Polymerase. The Journal of Biological Chemistry. 275(24): 18382-18390.
10. The Huntington's Disease Collaborative Research Group (1993) A novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes. Cell. 72: 971-983.
11. Jackson G, Salecker I, Dong X, Yao X, Arnheim N, Faber P, MacDonald M, Zipursly S (1998) Polyglutamine-Expanded Human Huntingtin Transgenes Induce Degeneration of *Drosophila* Photoreceptor Neurons. Neuron. 21: 633-642.
12. Johnson-Schlitz D, Flores C, Engels W (2007). Multiple-Pathway Analysis of Double-Strand Break Repair Mutations in *Drosophila*. PLoS Genetics. 3(4): e50.
13. Jung J, Bonini N (2007) CREB-Binding Protein Modulates Repeat Instability in a *Drosophila* Model for PolyQ Disease. Science. 315(5820): 1857-1859.
14. Jung J, van Jaarsveld M, Shieh S, Xu K, Bonini N (2011) Defining Genetics Factors That Modulate Intergenerational CAG Repeat Instability in Drosophila melanogaster. Genetics. 187(1): 61-71.
15. Kadyk LC, Hartwell LH (1992) Sister chromatids are preferred over homologs as substrates for recombinational repair in Saccharomyces cerevisiae. Genetics. 132(2):387-402.

16. Kane D, Shusterman M, Rong Y, McVey M (2012) Competition between Replicative and Translesion Polymerases during Homologous Recombination Repair in Drosophila. PLoS Genet. 8(4): e1002659.

17. Kovtun IV, Therneau TM, McMurray CT (2000) Gender of the embryo contributes to CAG instability in transgenic mice containing a Huntington's disease gene. Human Molecular Genetics. 9(18): 2767-2775.

18. Kovtun I, Liu Y, Bioras M, Klungland A, Wilson S, McMurray CT (2007) OGG1 initiates age-dependent CAG trinucleotide expansions in somatic cells. Nature. 477(7143): 477-452.

19. Kovtun I, McMurray CT (2008) Features of trinucleotide repeat instability *in vivo*. Cell Research. 18: 198-213.

20. Krench M, Littleton JT (2013) Modeling Huntington Disease in *Drosophila*. Fly. 7(4): 229-236.

21. Kuhl D, Caskey CT (1993) Trinucleotide repeats and genome variation. Curr. Opin. Genet. Devel. 3: 404-407.

22. Kunkel, TA (1993) Slippery DNA and diseases. Nature. 365: 207-208

23. La Spada A, Taylor J (2010) Repeat expansion disease: Progress and puzzles in disease pathogenesis. Nat Rev Genet. 11(4): 247-258.

24. Lenzmeier BA, Freudenreich CH (2003) Trinucleotide repeat instability: a hairpin curve at the crossroads of replication, recombination, and repair. Cytogenet Genome Res. 100:7-24.

25. Liu J, Heyer WD (2011). Who's who in human recombination: BRCA2 and RAD52. Proc Natl Acad Sci USA. 108(2): 441-442.

26. Macherey-Nagel (2014) Plasmid DNA purification User manual: Nucleospin Plasmid Easy Pure. Macherey-Nagel. Rev. 02.

27. McMurray, C (1999) DNA secondary structure: a common and causative factor for expansion in human disease. Proc. Natl. Acad. Sci. USA 96: 1823-1825.

28. McVey M, Lee S (2008) MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet. 24(11): 529-38.

29. McVey M (2010) In vivo analysis of Drosophila BLM helicase function during DNA double-strand gap repair. Methods Mol Biol. 587: 185-194.

30. Mirkin E, Mirkin S (2007) Replication Fork Stalling at Natural Impediments. Microbiol Mol Biol Rev. 71(1): 13-35.

31. Monteilhet C, Perrin A, Thierry A, Colleaux L, Dujon B (1990) Purification and characterization of the in vitro activity of I-sce I, a novel and highly specific endonuclease encoded by a group I intron. Nucleic Acids Research 18(6): 1407-1413.

32. Moynahan M, Jasin M (2010) Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. Nat Rev Mol Cell Biol. 11(3): 196-207.

33. Nimonkar AV, Sica RA, Kowalczykowski SC (2009) Rad52 promotes second-end DNA capture in double-strand break repair to form complement-stabilized joint molecules. Proc Natl Acad Sci USA. 106(9): 3077-3082.

34. Pâques F, Leung W, Haber JE (1998) Expansions and Contractions in a Tandem Repeat Induced by Double-Strand Break Repair. Molecular and Cellular Biology. 18(4): 2045-2054.

35. Pardo B, Gómez-González B, Aguilera A (2009) DNA double-strand break repair: how to fix a broken relationship. Cell Mol Life Sci. 66: 1039-56.

36. Petruska J, Arnheim N, Goodman M (1996) Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. Nucleic Acids Research. 24(11): 1992-1998.

37. Polleys EJ, House NCM, Freudenreich CH (2017) Role of recombination and replication fork restart in repeat instability. DNA Repair. 56: 156-165.

38. Preston C, Flores C, Engels W (2006) Differential Usage of Alternative Pathways of Double-Strand Break Repair in Drosophila. Genetics. 172(2): 1055-1068.

39. Pryor JM, Waters CA, Aza A, Asagoshi K, Strom C, Mieczkowski PA, Blanco L, Ramsden DA (2015) Essential role for polymerase specialization in cellular nonhomologous end joining. Proc Natl Acad Sci USA. 112(33): E4537-45.

40. Richard GF, Goellner GM, McMurray CT, Haber JE (2000) Recombination-induced CAG trinucleotide repeat expansions in yeast involve the MRE11-RAD50-XRS2 complex. The EMBO Joural. 19: 2381-2390.

41. Rodgers K, McVey M (2015) Error-Prone Repair of DNA Double-Strand Breaks. J Cell Physiol. 231: 15-24.

42. Rong YS, Golic K (2003) The homologous chromosome is an effective template for the repair of mitotic DNA double-strand breaks in Drosophila. Genetics 165(4): 1831-1842.

43. Roos R (2010) Huntington's disease: a clinical review. Orphanet Journal of Rare Diseases. 5(1): 40.

44. Sambrook J, Russell DW (2006) The Inoue method for preparation and transformation of competent E. coli: "ultra-competent" cells. Cold Spring Harb Protoc. 2:3944.

45. Shah K, Shishkin A, Voineagu I, Pavlov Y, Shcherbakova P, Mirkin SM (2012) Role of DNA polymerases in repeat-mediated genome instability. Cell Rep. 2(5): 1088-1095.

46. Steven NM (1908) A study of the germ cells of certain Diptera, with reference to the heterochromosomes and the phenomena of synapsis. J. Exp. Zool. 5: 359-383.

47. Stark J, Jasin M (2003) Extensive Loss of Heterozygosity Is Suppressed during Homologous Repair of Chromosomal Breaks. Mol Cell Biol 23(2): 733-743.

48. Sugawara N, Ira G, Haber J (2000) DNA Length Dependence on the Single-Strand Annealing Pathway and the Role of Saccharomyces cerevisiae RAD59 in Double-Strand Break Repair. Molecular and Cellular Biology 20(14): 5300-5309.

49. Toshiya Sato, Mutsuo Oyake, Kenji Nakamura, Kazuki Nakao, Yoshimitsu Fukusima, Osamu Onodera, Shuichi Igarashi, Hiroki Takano, Koki Kikugawa, Yoshinori Ishida, Takayoshi Shimohata, Reiji Koide, Takeshi Ikeuchi, Hajime Tanaka, Naonobu Futamura, Ryusuke Matsumura, Tetsuya Takayanagi, Fumiaki Tanaka, Gen Sobue, Osamu Komure, Mie Takahashi, Akira Sano, Yaeko Ichikawa, Jun Goto, Ichiro Kanazawa, Motoya Katsuki, Shoji Tsuji (1999) Transgenic Mice Harboring a Full-Length Human Mutant *DRPLA* Gene Exhibit Age-Dependent Intergenerational and Somatic Instabilities of CAG repeats Comparable with Those in DRPLA patients. *Human Molecular Genetics*. 8(1): 99–106.

50. Tran HT, Degtyareva NP, Koloteva NN, Sugino A, Masumoto H, Gordenin DA, Resnick MA (1995) Replication Slippage between Distant Short Repeats in *Saccharomyces cerevisiae* Depends on the Direction of Replication and the RAD50 and RAD52 Genes. Molecular and Cellular Biology. 15(10): 5607-5617.

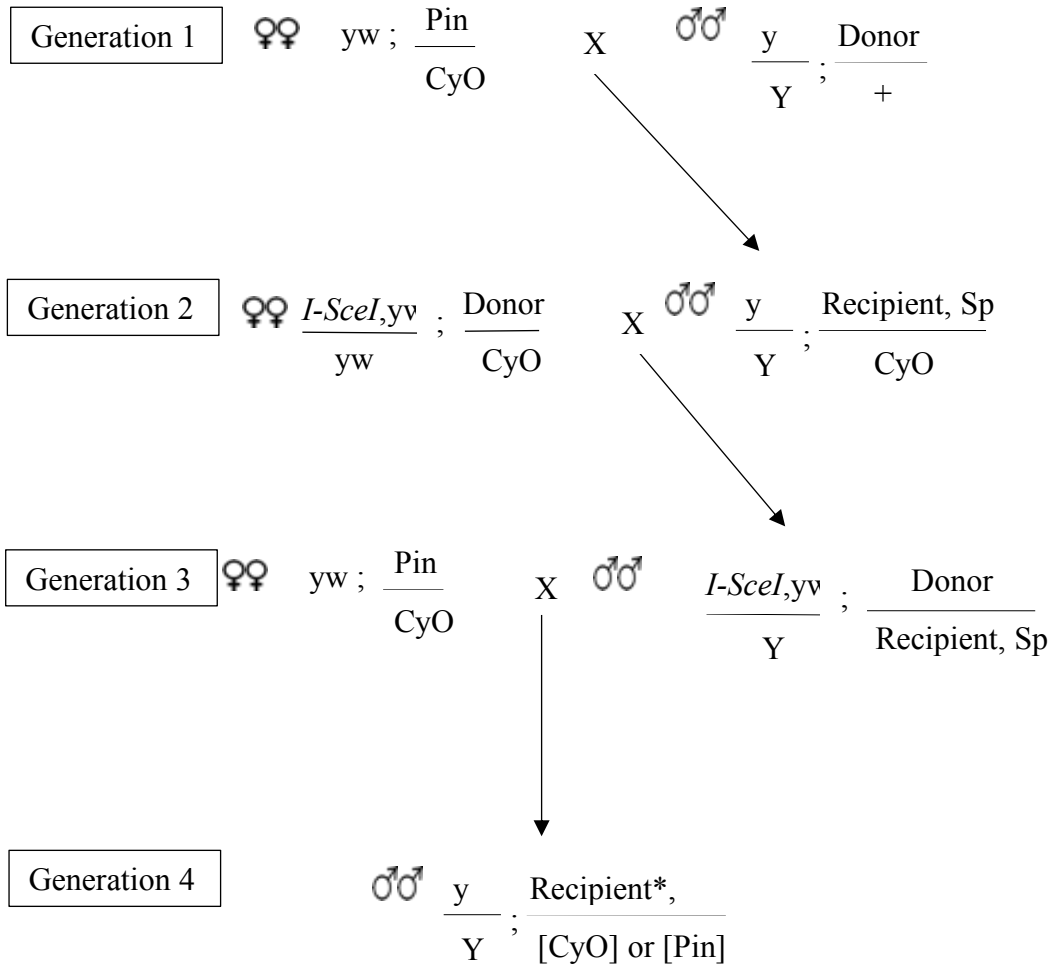
51. Viguera E, Canceill D, Ehrlich SD (2001) Replication slippage involves DNA polymerase pausing and dissociation. The EMBO Journal. 20(10): 2587-2595.

52. Vogelstein B, Gillespie D (1979) Preparative and analytical purification of DNA from agarose. Proc Natl Acad Sci USA 76(2): 615-619.

53. Yoon SR, Dubeau L, de Young M, Wexler NS, Arnheim N (2003) Huntington disease expansion mutations in humans can occur before meiosis is completed. Proc Natl Acad Sci USA. 100(15): 8834-8838.

54. Zhang X, Tee L, Wang X, Huang Q, Yang S (2015) Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Molecular Therapy – Nucleic Acids. 4, e264.

Supplemental Figures



Supplemental Figure 1. Fly cross to obtain generation 4 progeny with repair event.

Supplemental Table 1. Scoring sheet of fourth generation Sp males. Blacked out cells indicate vials that were scored incorrectly and therefore omitted during statistical analysis. Vials 6-12 and 30 also not included in analysis as the single males in that cross either did not contain the donor or recipient chromosome. Second and third counts not yet obtained for some vials.

Vial #	First count			Second count			Third count		
	End Joining (DsRed+)	Incomplete HR (DsRed-GFP-)	Complete HR (DsRed-GFP+)	End Joining (DsRed+)	Incomplete HR (DsRed-GFP-)	Complete HR (DsRed-GFP+)	End Joining (DsRed+)	Incomplete HR (DsRed-GFP-)	Complete HR (DsRed-GFP+)
1			0			1	0	0	1
2			0	11	0	0	19	0	0
3			1	16	0	0	7	0	0
4			1	7	0	0	3	0	1
5			0	5	1	0	4	0	0
13			0	1	0	0	0	0	0
14			0	3	0	0	5	0	1
15	6	2	0	5	0	2	3	0	3
16	12	0	0	4	0	0			
17	16	0	0	11	0	0			
18	15	0	0	4	0	0			
19	15	0	0	13	0	0			
20	10	0	0	13	0	0			
21	14	0	0	14	0	0			
22	16	0	0	5	0	0			
23	13	0	0	6	0	0			
24	5	0	1	12	0	0			
25	25	0	0	9	0	0			
26	10	0	1	9	0	0			
27	10	0	1	14	0	0			
28	14	0	0	4	1	0			
29	14	0	0	9	0	0			
31	6	0	0						
32	16	0	0						
33	13	0	0						
34	16	0	1						
35	11	0	0						
36	17	0	0						
37	9	0	0						
38	8	0	0						

Supplemental Table 2. Repair pathways utilized via differential positioning of *I-SceI*. UIE-5B used in our assay. (1) represents a cross in which the homolog had no template for repair, while (2) represents data from the cross in which a template in present on the homolog for repair. Data from Preston et al., 2006.

Endonuclease source	% UIE-72C[a]	% UIE-5B[b]	% UIE-87F[c]	% UIE-2R[d]	% UIE-53D[e]
SSA (1)	68.4 ± 1.7	76.7 ± 1.3	62.5 ± 1.1	74.2 ± 1.0	73.0 ± 1.3
NHEJ (1)	31.1 ± 1.3	24.9 ± 1.7	34.6 ± 1.1	18.6 ± 0.6	25.2 ± 1.1
Total (1)	99.5	101.6	97.1	92.8	98.2
SSA (2)	66.1 ± 1.6	61.3 ± 1.6	74.5 ± 1.6		
NHEJ (2)	18.7 ± 2.1	17.5 ± 1.0	16.2 ± 1.1		
Conversion	11.3 ± 1.6	19.4 ± 1.2	7.7 ± 1.3		
Total (2)	96.1	98.2	98.4		