Detection and Analysis of Common Fragile Sites in *Drosophila melanogaster*

Matthew C. LaFave

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Genetics & Molecular Biology.

Chapel Hill
2011

Approved by:

Jeff Sekelsky

Shawn Ahmed

Corbin Jones

Steve Matson

Dale Ramsden

Abstract

Matthew C. LaFave: Detection and Analysis of Common Fragile Sites in *Drosophila melanogaster*
(Under the direction of Jeff Sekelsky)

Common fragile sites (CFSs) are regions of DNA exceptionally prone to breakage. While these regions have implications in cancer, the causes of chromosome fragility remain poorly understood. This is partially due to relatively low-resolution cytological mapping of CFSs, and the use of exogenous agents to induce chromosome breakage. In an effort to better understand the causes of fragility, I have developed *Drosophila melanogaster* as a model for CFS study. In doing so, I have developed approaches to identify CFSs at a high resolution, based on both spontaneous chromosomal events and breakage induced by the inhibition of replication. In the first approach, I used a mutant form of *mus309*, the ortholog of human BLM helicase, to locate CFSs by visualizing sites of DNA breakage as mitotic crossovers. High rates of breakage correspond to CFSs, and results from my study indicate that there is a significantly non-uniform rate of mitotic crossovers across the left arm of chromosome *2*. This work constitutes the first report of specific regions sensitive to endogenous damage in *D. melanogaster*. Further, the resolution of damage detection can be brought even higher with SNP mapping. My second approach is a novel assay that uses S2 cell culture to detect preferential sites of exogenous DNA integration, a hallmark of CFSs. This allows me to survey the entire genome, while using high-throughput sequencing to obtain a resolution of CFS detection orders of magnitude better than previous studies. I obtained numerous integration events under a variety of conditions, providing evidence of putative fragile regions. I anticipate that the assays I developed will serve as valuable tools for the detection of DNA breakage in future studies. Further, I expect the characterization

of the CFSs detected from both of these approaches will lead to a better understanding of the causes of inherent genome instability.

To Kelly, who will always be my Player 2.

Acknowledgements

I would first like to thank my advisor, Dr. Jeff Sekelsky. He has been unfailingly supportive and encouraging, and has made graduate school something to be excited about. I thank the members of my committee, Dr. Shawn Ahmed, Dr. Corbin Jones, Dr. Steve Matson, and Dr. Dale Ramsden, for their helpful advice pertaining to both my experiments and my career. I thank the members of the Sekelsky lab, for making the lab an ideal work environment. I'd like to give particular recognition to Jeannine LaRocque, who mentored me as a rotation student, and Susan McMahan Cheek, whose appreciation of Ratt was one of the first signs that the Sekelsky lab was the right fit. I'd like to recognize all of the Sekelsky lab undergrads, as well, but particularly my mentees Lewis Overton and Kelly Wolfe. Both intelligent and hard-working, they endured my first forays into mentoring, in which I learned at least as much as I taught them. I'd also like to thank Dr. Bob Duronio and his lab for helpful discussions in lab meetings. I'd like to specifically thank Bob for our meetings when I was in my first year, when he confirmed that, yes, it would be a good idea to do a rotation in the Sekelsky lab. I'm grateful to the UNC Center for Bioinformatics, particularly Dr. Hemant Kelkar and Dr. Xiaojun Guan. The assistance of Dr. Piotr Mieczkowski and the UNC High-Throughput Sequencing Facility was instrumental in this project, as well. I'd like to thank some of my influential professors from my time as an undergrad at Notre Dame: Dr. Michelle Whaley, Dr. Gary Lamberti, Dr. Paul Helquist, and Dr. William Ramsey. Among my undergrad professors, I'd most like to thank my first research advisor, Dr. David Hyde, as well as Dr. Christopher Burket, the postdoc who served as my mentor. I'd also like to thank Doug Baltz, my excellent high school physics teacher. I acknowledge my sources of funding: the GMB training grant, the CMB training grant, and the Office of Undergraduate

Research. Thanks to the friends who have made my time at UNC so much more than just work, especially my comrades-in-arms in Team Venture. I'd like to recognize the motivational prowess of Steve Harris, Bruce Dickinson, Kai Hansen, Sam Totman, and ZP Theart. Of course, I thank my family, who has supported me in everything I've done. My mother's use of a Punnett square to explain why I was the only one in the family with blue eyes encouraged me to take my first step into a larger world. Finally, I thank my fiancée, Kelly. She is the most loving and giving person I have ever known, and the very embodiment of patience.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BAC – Bacterial artificial chromosome

BrdU – Bromodeoxyuridine

CFS – Common fragile site

ChIP-seq – Chromatin immunoprecipitation followed by high-throughput sequencing

CO – Crossover

DmBLM – *Drosophila melanogaster* ortholog of BLM (encoded by *mus309*)

DMSO – Dimethyl sulfoxide

DSB – Double-strand break

FHA – Forkhead-associated

GA IIx – Genome analyzer IIx

HTS – High-throughput sequencing

Polα – Polymerase alpha (encoded by *DNApol-α180*)

RACE – Rapid amplification of cDNA ends

RFS – Rare fragile site

ssDNA – single-stranded DNA

UTR – Untranslated region

Chapter I

General Introduction

Genome stability & fragile sites

DNA serves as the master plan for the cell, encoding the information for processes necessary to begin and sustain life. Just as the information encoded by the genome is important, the physical integrity of the DNA is critical to ensure successful implementation of the coded information. When genome stability is compromised, the negative effects on the cell or organism can be varied and severe (DILLON *et al.* 2010). It is critical, therefore, for the cell to maintain the integrity of its genome. The cell has evolved numerous mechanisms to cope with this issue (SANCAR *et al.* 2004). However, damage can occur, and often does. Damage can come from exogenous sources, such as UV radiation or chemicals, or endogenous sources, such as byproducts of cellular metabolism or errors during DNA replication (DE BONT and VAN LAREBEKE 2004). Interestingly, some regions of DNA appear to be more prone to damage than others.

Rare fragile sites

Regions of DNA that are exceptionally prone to chromosomal abnormalities – chiefly breakage – have been termed “fragile sites” (DURKIN and GLOVER 2007). When a fragile site forms a break or gap, it is said to be expressed. Fragile sites have two classifications: rare and common. Rare fragile sites (RFSs) are found infrequently in the human population, and are associated with specific disease alleles (SUTHERLAND *et al.* 1998). Most RFSs are folate-sensitive, meaning that they express their fragile characteristics in cells cultured in folate-deficient media (LUKUSA and FRYNS 2008). The few RFSs that are not folate-sensitive have their expression induced by bromodeoxyuridine (BrdU) and/or distamycin A, agents that integrate into DNA and interfere with replication (LUKUSA and FRYNS 2008).

RFS only exhibit fragility in an individual with the disease. For example, the cultured cells of an individual with Fragile X disease would exhibit a chromosome constriction at the FRAXA RFS, whereas those of a healthy individual would not (MCBRIDE 1979; VERKERK *et al.* 1991). The causes of RFS fragility are relatively well understood: the expansion of di- or trinucleotide repeats leads to non-B DNA structures, which interfere with DNA replication and nucleosome assembly. The wild type copy number of such repeats is tolerated by the cell, but expansion beyond a threshold value specific to each RFS leads to chromosomal instability (LUKUSA and FRYNS 2008). The increased prevalence of non-B DNA makes these regions prone to DNA gaps and breaks (FREUDENREICH 2005).

Common fragile sites

Common fragile sites (CFSs), on the other hand, are generally considered to be a normal component of chromosome structure. There have been 88 CFSs identified in the human genome, thirteen of which have been characterized at a resolution higher than that of chromosome bands (LUKUSA and FRYNS 2008). CFSs are presumed to be ubiquitous in the human population, although some studies have indicated variability in the fragility of sites between individuals, and between different tissues (DENISON *et al.* 2003; TEDESCHI *et al.* 1992). In addition, CFSs are not all equally fragile; the first study of CFSs found that, in humans, 80% of breaks at CFSs were due to only 20 CFSs (GLOVER *et al.* 1984). The CFSs FRA3B and FRA16D are the most highly expressed.

Human CFSs have a medically relevant connection to cancer. When expressed in the presence of aphidicolin, CFSs have been shown to serve as significant predictor of breast, ovarian, and lung cancer (DHILLON *et al.* 2003). Breaks at CFSs are associated with genomic rearrangements implicated in cancer, including loss of heterozygosity (DURKIN and GLOVER 2007). Recurrent breaks at CFSs have been found in prostate, breast, lung, and pancreatic cancer, among many other types (CARTER *et al.* 1990; CHEN *et al.* 1996; HIBI *et al.* 1992; NEGRINI *et al.* 1996; SHRIDHAR *et al.* 1996). In fact, most cancer-specific translocations have at least one breakpoint in a fragile site (BURROW *et al.* 2009). Furthermore, the human CFSs most prone to breakage, FRA3B and FRA16D, lie within tumor suppressor genes (HUEBNER and CROCE 2003; O'KEEFE and RICHARDS 2006). This indicates that breaks in these regions, especially if they are repaired incorrectly or go unrepaired, could contribute to downstream events involved in tumorigenesis. Indeed, breaks at CFSs have been shown to be capable of initiating gene amplification via breakage-fusion bridge cycles (COQUELLE *et al.* 1997).

The strict definition of what constitutes a CFS comes from the way they are traditionally detected. CFSs are defined as choromsomal regions that form reproducible breaks on metaphase chromosomes after partial inhibition of DNA replication (DURKIN and GLOVER 2007). Breaks can be observed in cells grown in folate-deficient media or in the presence of fluorodeoxyuridine, although the level of breakage is low (GLOVER 1981; YUNIS and SORENG 1984). Efficient inhibition of replication is therefore usually achieved by use of aphidicolin, and occasionally by agents such as BrdU or 5-azacytidine (GLOVER *et al.* 1984; SUTHERLAND *et al.* 1985). Aphidicolin acts to inhibit polymerases involved in DNA replication, such as polymerase α, δ, and ε (CHENG and KUCHTA 1993; GOSCIN and BYRNES 1982; IKEGAMI *et al.* 1978). While aphidicolin can be used to completely halt the cell cycle in S phase, CFSs are induced using a low concentration of the agent, typically 0.4 μM (GLOVER *et al.* 1984). At this concentration, aphidicolin makes the replication of DNA somewhat more difficult, without causing it to stop altogether.

It has been observed that down-regulation of components of homologous recombination, such as RAD51, or of non-homologous end joining, such as DNA-PKcs, results in a significant increase in CFS expression in cells treated with aphidicolin (SCHWARTZ *et al.* 2005). This indicates that expressed CFSs can be repaired both by homologous repair and non-homologous end-joining. CFSs have been reported to be prone to sister chromatid exchanges after treatment with aphidicolin, a result of homologous repair of the double-strand break (DSB) incurred at the CFS (GLOVER and STEIN 1987; HIRSCH 1991). Of course, not all expressed CFSs are repaired after exposure to replicative stress, leading to the characteristic breaks on metaphase chromosomes.

CFSs appear to be evolutionarily conserved

An interesting, and somewhat counter-intuitive, characteristic of CFSs is that they appear to be evolutionarily conserved. Most studies of CFSs are done with human chromosomes, but CFSs have also been detected in several other mammals. It has been shown that mice have CFSs that correspond to known human CFSs (HELMRICH *et al.* 2006; HELMRICH *et al.* 2007). These studies in mice showed conservation of specific CFSs, but regions analogous to CFSs – that is, regions prone to breakage, especially when under replicative stress – have also been detected in yeast.

Reducing levels of polymerase α in *Saccharomyces cerevisiae* has been shown to increase the rate of breaks leading to translocations, due in large part to retrotransposons called Ty elements (LEMOINE *et al.* 2005). A site at which Ty elements formed an inverted repeat was found to be prone to double-strand breaks under these conditions; therefore, it has the characteristics of a CFS. A separate study found that a region containing tRNA genes was prone to DNA breaks (ADMIRE *et al.* 2006). Disruption of replication by exogenous agents or mutation of a helicase resulted in increased instability in this region, and removal of the region reduced the overall genomic instability of the cell. Finally, a study of spontaneous mitotic recombination in yeast identified a 1.3 kb region that was significantly more prone to crossovers than surrounding regions (LEE *et al.* 2009). We published a Perspectives article on this work, in which we pointed out that such a region might constitute a CFS prone to endogenous damage (LAFAVE and SEKELSKY 2009).

Proposed causes of CFS fragility

It is important to note that the reasons CFSs are prone to damage are not well understood. CFSs do not appear to share the nucleotide repeats characteristic of RFSs, so the reason for CFS fragility is likely to be somewhat different (ARLT *et al.* 2002; RASSOOL *et al.* 1996; SCHWARTZ *et al.* 2006). CFSs can be induced by agents that interfere with replication, so it is generally thought that some aspect of fragile regions that make them inherently difficult to replicate.

This claim is bolstered by genetic evidence, which involves the checkpoint kinase Ataxia-telengiectasia and Rad3 Related (ATR). This protein is involved in signaling the response to stalled or collapsed replication forks, as well as the response to single stranded DNA exposed by damage. In human cell culture, ATR has been shown to be necessary for CFS stability (CASPER *et al.* 2002). CFS expression is enhanced in the presence of aphidicolin in *ATR* mutant cells, and such cells spontaneously express CFSs even in the absence of aphidicolin. ATR has been found to preferentially interact with FRA3B after treatment with aphidicolin (WAN *et al.* 2010). Depletion of CHK1, the downstream effector of ATR, also results in CFS expression (DURKIN *et al.* 2006). Similar results have been reported in yeast, in which cells mutant for the ortholog of ATR, *Mec1*, display an increase in DNA breaks in replication slow zones (CHA and KLECKNER 2002). These reports suggest that a key early step in CFS expression is the stalling of a replication fork, consistent with the notion that CFSs are inherently difficult to replicate. What causes the forks to stall, however, remains an open question. Here, I discuss some of the major properties that have been proposed to contribute to CFS fragility.

Primary sequence characteristics

Many characteristics of CFSs have been proposed as the reasons for their fragility. Primary sequence attributes have received a considerable amount of attention in this regard. One model posits that aphidicolin causes uncoupling of polymerases from the helicase-topoisomerase complex at the replication fork, resulting in exposed single stranded DNA (ssDNA) (DURKIN and GLOVER 2007). Some of the sequences in CFSs have the potential to form stable secondary structures – therefore, exposure of such sequences as ssDNA may lead to hairpins and other structures that disrupt replication. Interestingly, the breaks and ssDNA at CFSs can be reduced by low-dose camptothecin, even in the presence of aphidicolin (ARLT and GLOVER 2010). It is thought that low concentrations of camptothecin may slow the helicase-topoisomerase complex, and that this may reduce the polymerase uncoupling normally induced by aphidicolin.

No sequence motifs have been detected in CFSs, although it is unclear if the relatively low resolution of CFS detection makes such regions difficult to identify (MISHMAR *et al.* 1998; SCHWARTZ *et al.* 2006). Still, there is evidence that the sequence of CFSs is inherently unstable. It was shown in a study in which bacterial artificial chromosomes (BACs) carrying sequence from FRA3B were inserted at ectopic sites of the human genome that the FRA3B BACs could recapitulate fragility, while control BACs did not (RAGLAND *et al.* 2008).

In addition, DNA flexibility has been proposed to be a major cause of CFS fragility (SCHWARTZ *et al.* 2006). Flexibility calculations are based on the maximum potential angular twist that two consecutive bases could achieve (SARAI *et al.* 1989). A program, TwistFlex, was developed to calculate the average potential twist of DNA in a sliding window; averages that surpass a threshold value are reported as flexibility peaks (MISHMAR *et al.* 1998). Several CFSs have been analyzed in this way, and found to have a high number of flexibility peaks (DURKIN and GLOVER 2007). Interestingly, a highly flexible region of human FRA16D inserted into the *S. cerevisiae* genome was shown to induce fragility, while non-flexible regions did not (ZHANG and FREUDENREICH 2007). The reported flexibility of such regions is likely a result of the AT-rich nature of CFSs, as the A to T step is more than twice as flexible as any other possible step (SARAI *et al.* 1989). These regions could contribute to fragility *via* secondary structure formed by AT repeats. However, if DNA flexibility is involved in CFS fragility, it is unlikely to be the sole cause. Not all CFSs have more flexibility peaks than controls, and BACs with a high number of flexibility peaks are not sufficient to recapitulate fragility in ectopic regions (HELMRICH *et al.* 2007; RAGLAND *et al.* 2008).

Chromatin environment

The state of the chromatin surrounding CFSs has also been considered to contribute to fragility. Though not strictly related to replication problems caused by aphidicolin, the banding pattern of metaphase chromosomes has raised an interesting issue regarding fragility. The dark G-bands tend to be AT-rich, late replicating, and have few genes; the light R-bands are just the opposite: early-replicating and gene- and GC-rich (GARDINER 1995). Some CFSs, including FRA3B, have the characteristics of the dark G-bands, but map to R-bands. It has been suggested that CFS fragility may be related to the discrepancy between the chromatin context of CFSs and their surrounding environment (SCHWARTZ *et al.* 2006). Others have suggested that histone modifications may play a role in fragility. Histones at CFSs have been found to be hypoacetylated, and FRA3B is more resistant to micrococcal nuclease than nearby non-fragile regions (JIANG *et al.* 2009). These findings suggest that CFS chromatin might be more compact than flanking regions, and raise questions about the role of chromatin in fragility.

Replication timing & origin density

The timing of replication, and the related factor of the distance from a firing replication origin, may have a substantial impact on fragility. CFSs are often associated with large, late-replicating genes in humans (LE BEAU *et al.* 1998). It has been reasoned that these regions are fragile because they are unable to replicate their DNA by the end of S phase, especially in the presence of replication-inhibiting aphidicolin. FRA16D, for example, was found to have a slow-moving replication fork (PALAKODETI *et al.* 2004). Recently, it was determined that the center of FRA3B is origin-poor in JEFF lymphocytes (LETESSIER *et al.* 2011). Moreover, in cells in which FRA3B was not origin-poor, such as MRC-5 fibroblasts, fragility was not observed. This suggests that the combination of replication timing and origin density play a major role in CFS fragility, and may explain tissue-specific differences in CFS expression. However, studies that have found that insertions of CFS DNA in ectopic regions are sufficient to recapitulate fragility, while control insertions are not, suggest that origin density is insufficient to fully explain fragility (RAGLAND *et al.* 2008; ZHANG and FREUDENREICH 2007).

Issues of resolution

While much has been learned about the characteristics of CFSs, studies of CFSs have suffered from the use of relatively low-resolution approaches. The initial mapping of CFSs were accomplished by observing breaks on metaphase chromosome spreads, and were thus limited to the resolution of chromosome bands (GLOVER *et al.* 1984). Fluorescent *in situ* hybridization probes have been used to achieve a somewhat higher resolution (BECKER *et al.* 2002). These studies score breaks qualitatively – proximal, distal, or within – relative to a probe of known location. By doing this for many probes, a distribution of breaks can be generated. This technique is good for determining the outer boundaries of fragility and general shape of the distribution, but the effective resolution is still in the hundreds of kilobases. This is because the precise location of any given break cannot be inferred from its position relative to a probe. Because of this, CFSs have been classified as large, fragile regions, often a megabase or more in length (BECKER *et al.* 2002; HANDT *et al.* 2000).

This resolution makes it difficult to ascertain which factors truly contribute to fragility. For example, it leaves us unable to distinguish between a single large region of fragility, and several smaller regions of fragility that cluster together. To answer the question of the nature of fragility, it would be advantageous to focus on what I will refer to as the minimal fragile region – the most fragile portion of the CFS. Including in the analysis regions that are less fragile may have the effect of drowning out the truly relevant information from the minimal fragile region.

Induced vs. natural CFSs

Almost all of what we know about CFSs is based on damage induced by exogenous sources, typically aphidicolin. The use of such inducers of fragility has been necessitated by the difficulty of locating sites of endogenous damage in human cell culture. The physiological relevance of these aphidicolin-induced regions is well-established; the most fragile CFSs, such as FRA3B and FRA16D, have been implicated in recurrent tumorigenic breakpoints (HUEBNER and CROCE 2003; O'KEEFE and RICHARDS 2006). Still, I felt it would be useful to determine the location of CFS breaks caused by endogenous lesions. This constitutes DNA damage that the genome normally incurs, but typically repairs in a way that makes the initial break difficult to detect.   I designed assays in a way that allowed me to examine damage incurred under both types of conditions. I will refer to CFSs corresponding to endogenous breaks as “natural CFSs”; those that have their fragility induced by exogenous sources will be referred to as “induced CFSs”. By analyzing both types of regions, one can determine if induced CFSs are exceptionally aphidicolin-sensitive, or if aphidicolin simply increases the expression of natural CFSs.

Using *Drosophila* to study CFSs

To accomplish the tasks of studying CFSs at a high-resolution, induced by endogenous or exogenous means, I chose to work with *Drosophila melanogaster*. To my knowledge, there have been no previous studies to detect CFSs in *Drosophila*. Most studies of CFSs have taken place in human cell culture, and a few have used yeast; this study gives me an opportunity to develop *D. melanogaster* as the first live, whole-organism *in vivo* metazoan model of CFSs. There exists an extensive genetic toolkit for *Drosophila*, with many useful mutations available, not the least of which is a mutation that greatly increases the frequency of mitotic crossovers (MCVEY *et al.* 2007). My implementation of an assay that takes advantage of the characteristics of this mutant to detect endogenous damage and verify the presence of CFSs in *D. melanogaster* is detailed in Chapter II. This is the first report of CFSs detected in *Drosophila*. Having established the presence of such sites, I developed a cell-based assay that couples selectable DNA integration with high-throughput sequencing (HTS) to identify putative CFSs at a high resolution, and on a genome-wide scale. The assay, as well as my analysis of the integration sites, is detailed in Chapter III. I have analyzed the results of these assays to offer insight into the nature of fragility.

Chapter II

Common Fragile Sites and Mitotic Crossovers

Introduction

Common fragile sites (CFSs) are chromosomal regions that are prone to breakage, particularly when under DNA replication stress. While the existence and location of human CFSs are well-documented (DURKIN and GLOVER 2007), the relatively low resolution at which CFSs have been studied means that an understanding of the causes of fragility has remained elusive. I aimed to determine the causes of fragility by employing high-resolution approaches in *Drosophila melanogaster*. By increasing the resolution at which CFSs are detected and analyzed, we increase the chance of determining features unique to CFSs. In addition, it permits one to distinguish between a CFS as a single, large region of fragility, or several smaller regions clustered close together.

This study established *D. melanogaster* as a metazoan *in vivo* model of CFSs. This particular combination has not been available to the CFS field before – studies are typically carried out in human cell culture (GLOVER *et al.* 1984), and occasionally in sacrificed mice (HELMRICH *et al.* 2006). So far, the only truly *in vivo* studies of CFSs have been in budding yeast (ADMIRE *et al.* 2006; LEMOINE *et al.* 2005). Working in *Drosophila* allows us to bridge studies in yeast and mammals, and to ask questions about the evolutionary conservation of CFSs.

The first step was to determine the location of *D. melanogaster* CFSs – indeed, to see if flies have CFSs at all. Data in yeast, as well non-human mammals such as mice, dogs, and cats, suggest that such regions are conserved (HELMRICH *et al.* 2007; LEMOINE *et al.* 2005; STONE *et* *al.* 1991; STONE *et al.* 1993). However, no one had looked for CFSs in *Drosophila*, and so detection of the sites became my primary task.

I chose to use a novel genetic assay to determine the presence and location of CFSs (Fig. 2.1). My assay allowed me to map sites of DNA breakage by taking advantage of a characteristic of flies mutant for *mus309*, which encodes *Drosophila* BLM (DmBLM). DmBLM is the *Drosophila* homolog of human BLM, a RecQ helicase (ADAMS *et al.* 2003; BACHRATI and HICKSON 2003). Humans that lack BLM have Bloom's Syndrome (BS), characterized by short stature, sterility, sensitivity to sunlight, and increased susceptibility to a broad spectrum of cancers. This last phenotype underscores the role of BLM as a fundamental protein in DNA repair. BS cells display an increased rate of sister chromatid exchanges (CHAGANTI *et al.* 1974). BLM can act on structures formed during homologous repair, such as migrating Holliday junctions and unwinding D-loops *in vitro* (KAROW *et al.* 2000). BLM also acts to maintain the integrity of stalled replication forks (DAVIES *et al.* 2007; MANKOURI and HICKSON 2007). In *Drosophila*, there is evidence suggesting that DmBLM acts to free newly-synthesized DNA from the template D-loop during homologous repair (MCVEY *et al.* 2004).

The feature of *mus309* mutants most relevant to this study is the greatly increased rate of mitotic crossovers (MCVEY *et al.* 2007). Such crossovers are associated with sites of endogenous breaks (Fig. 2.1). In flies with wild type *mus309*, DNA damage in mitotically dividing cells is typically repaired to yield a non-crossover product. Mitotic crossover frequencies in this situation are typically below 0.002% between *dp* and *bw*, a region covering a little under 30% of the genome (WOODRUFF and THOMPSON 1977). In flies homozygous mutant for *mus309*, however, mitotic crossover frequencies are about 2% between *st* and *e*, a region that covers a little over 20% of the genome (MCVEY *et al.* 2007). I designed an assay in which these crossovers would occur in the male germline, so the location of the crossover – and therefore the initial site of endogenous damage – could be detected via mapping of recessive markers in the progeny of the male.

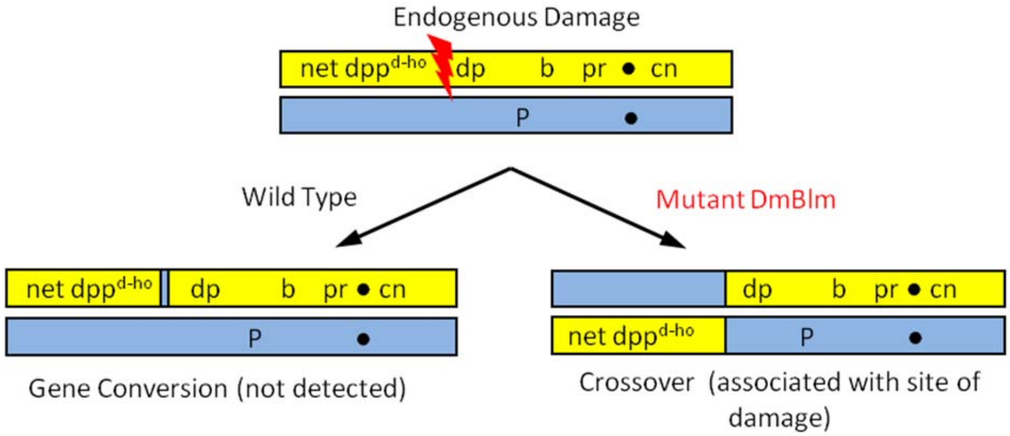

Endogenous Damage
net dpp^{d-ho} dp b pr ● cn
P ●
Wild Type
Mutant DmBlm
net dpp^{d-ho} dp b pr ● cn
P ●
Gene Conversion (not detected)
dp b pr ● cn
net dpp^{d-ho} P ●
Crossover (associated with site of damage)

Figure 2.1. Detection of fragile sites. Genomic damage incurred in mitotic cells of wild-type flies is repaired to yield a non-crossover product, making the site of damage difficult to detect. The *mus309* mutant flies lack DmBLM protein, which results in damage being repaired to produce a mitotic crossover product. If this occurs in the male germline, it will yield progeny in which the location of the crossover, and therefore the site of damage, can be mapped. The marker genes and P-element used for mapping are represented on the chromosomes; the ● indicates the centromere.

A key feature of the assay is that it allows one to map breaks incurred from endogenous sources. Although such damage, which cells normally incur, is the most biologically relevant, most studies of CFSs have focused on damage induced in the presence of exogenous agents, usually aphidicolin (DURKIN and GLOVER 2007). This is because endogenous damage normally occurs at a frequency that is difficult to detect with traditional, cytology-based approaches to mapping CFSs (GLOVER *et al.* 1984). My assay detects endogenous damage at a sufficiently high frequency to make the analysis of such damage straightforward. This permits me to determine the location of natural CFSs.

In addition, the assay can be modified to detect breakage in the presence of replicative stress. Direct treatment with aphidicolin is possible, but would be impractical – flies would have to be fed the chemical, and it would be very difficult to ensure that each fly consumed exactly the same dosage. However, I was able to create a mutant that genetically mimics the effect of aphidicolin. Since aphidicolin acts to inhibit replicative polymerases, I removed one copy of *DNApol-α180,* the gene encoding for the catalytic subunit of polymerase α (LAROCQUE *et al.*

2007a); *DNApol-α180* will henceforth be referred to as *Polα*). In this way, the assay can be used to detect either endogenous damage, or damage like that induced by aphidicolin.

The assay is also able to recover both halves of a DNA damage repair event. This is noteworthy, because one typically is only able to recover and examine one half of a mitotic crossover in metazoans, limiting the amount of information that can be gleaned from such a study. In my assay, mitotic crossovers occur pre-meiotically in the male germline, meaning that both halves of the crossover are produced as gametes. Analysis of the repair event from reciprocal crossover siblings gives us the ability to determine the nature of repair in a *mus309* mutant background.

I employed this assay to determine if CFSs are present on the left arm of chromosome *2* in *D. melanogaster*. Here, I treat CFSs as regions that are prone to DNA breakage; the source of this damage can be either endogenous or the result of replication inhibition. My assay allows one to distinguish between the two sources of damage, and lends itself to insights regarding the nature of chromosome fragility. I argue that *D. melanogaster* does contain CFSs, both induced and natural, and identify fragile regions on *2L*.

Results

It was first necessary to determine if the fly genome contains regions that are especially prone to DNA breakage. I accomplished this through the use of a mitotic crossover assay (Fig. 2.2). In these *mus309* mutants, sites of DNA breakage can be repaired to produce a mitotic crossover. The crossover is located at the site of initial damage, so by using phenotypic markers to map a distribution of crossovers, I was able to essentially generate a visualization of the distribution of breakage.

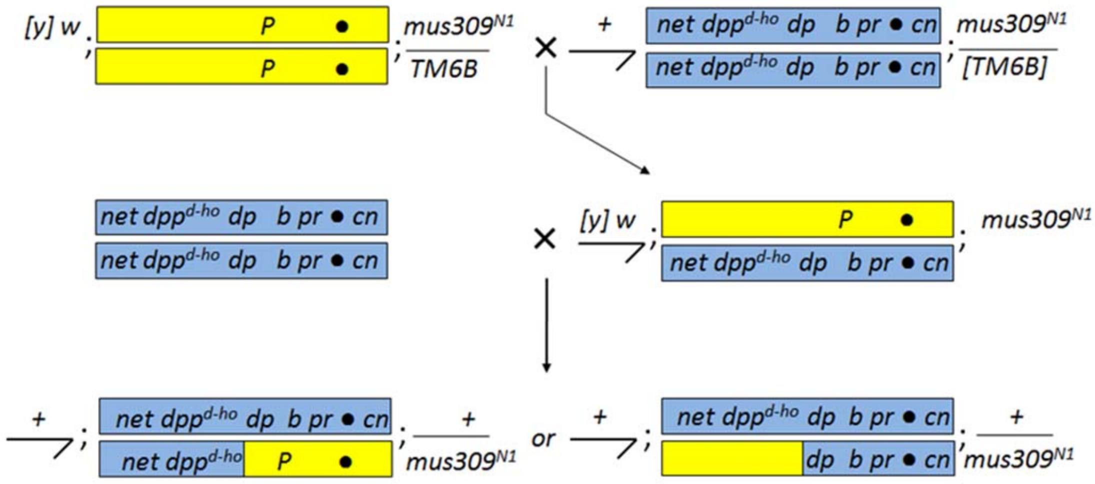

[y] w
P ●
P ●
mus309^N1
TM6B
×
+
net dpp^d-ho dp b pr ● cn
net dpp^d-ho dp b pr ● cn
mus309^N1
[TM6B]
net dpp^d-ho dp b pr ● cn
net dpp^d-ho dp b pr ● cn
×
[y] w
P ●
net dpp^d-ho dp b pr ● cn
mus309^N1
+
net dpp^d-ho dp b pr ● cn
net dpp^d-ho P ●
+
mus309^N1
or
+
net dpp^d-ho dp b pr ● cn
dp b pr ● cn
+
mus309^N1

Figure 2.2. Cross scheme to visualize mitotic crossovers. Virgin females heterozygous for *mus309*^{N1} and homozygous for a *P* element on *2L* were crossed to *mus309*^{N1} males that carried a 2nd chromosome with six recessive phenotypic markers (Cross 1). From the progeny of this cross, males that did not carry *TM6B*, and were therefore homozygous for *mus309*^{N1}, were collected. These flies carried one copy of each of the parental chromosomes, and the crossovers I was able to analyze in the following generation were produced in the pre-meiotic germline of these males. Each male was crossed to three virgin females homozygous for the marker chromosome (Cross 2). The progeny of this cross were scored for mitotic crossovers. One or both halves of the crossover could be recovered, and the *P* element could be detected via PCR to serve as an additional marker. The ● indicates the centromere.

In the first iteration of the assay, I analyzed 532 crossovers from 313 *mus309*^{N1} males. By examining the distribution of mitotic crossovers on *2L*, I was able to determine that the rate of crossing over, and therefore the rate of endogenous damage, is significantly non-uniform (bootstrapping, $P < 0.0001$; Fig. 2.3). This suggests that regions that produced the highest rates of crossing over – specifically, the regions between *net* and *dp*, and between *b* and *pr* – constitute CFSs, or that they contain more CFSs than the other intervals. In release 5 of the *D. melanogaster* genome, these regions cover the regions *2L*:87,382..4,479,471 and *2L*:13,823,894..20,073,719, respectively. These results were recapitulated in additional experiments in which stocks with different markers were used to derive the 2nd chromosome homologs, and where a different combination of *mus309* alleles was used ($P = 0.0002$; Figs. 2.4 and 2.5). In this version of the assay, I analyzed 634 crossovers from 391 *mus309*^{N1}/*mus309*^{D2}

males. The heteroallelic combination of *mus309* alleles demonstrates that the crossovers detected in the assay are due to the lack of functional DmBLM, and not to homozygous second-site mutations. The *al-dp* fragile region detected in this version of the assay covers the region *2L*: 387,439..4,479,471.

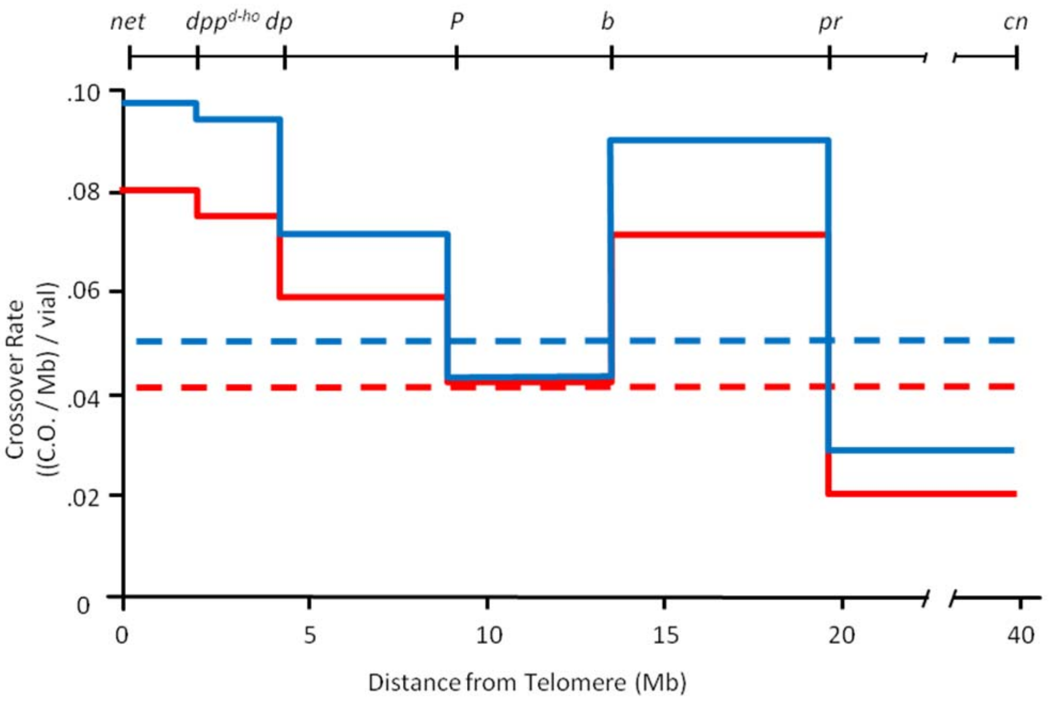


Figure 2.3. The rate of crossovers is non-uniform. Five phenotypic markers on *2L* (*net-pr*) and one on *2R* (*cn*) were used to determine the rate of mitotic crossovers in the male germline. One homolog of chromosome *2* carried the markers, and the other carried a *P* element to increase the resolution of the 9.2 Mb between *dp* and *b*. PCR was used to detect the *P* element. Crossover data was collected in vials of three females crossed to a single male; each vial therefore represents crossovers in the germline of one male. The red line represents males homozygous mutant for *mus309*^{N1} (532 crossovers, 313 vials), while the blue line represents males that were both homozygous mutant for *mus309*^{N1} and heterozygous mutant for *DNApol-α180*, the gene for Polα (334 crossovers, 157 vials). Dotted lines are the average crossover rates over the entire 40.6 Mb interval. The region between *pr* and *cn* contains about 16.4 Mb of heterochromatin, which is included in the calculation of the crossover rate.

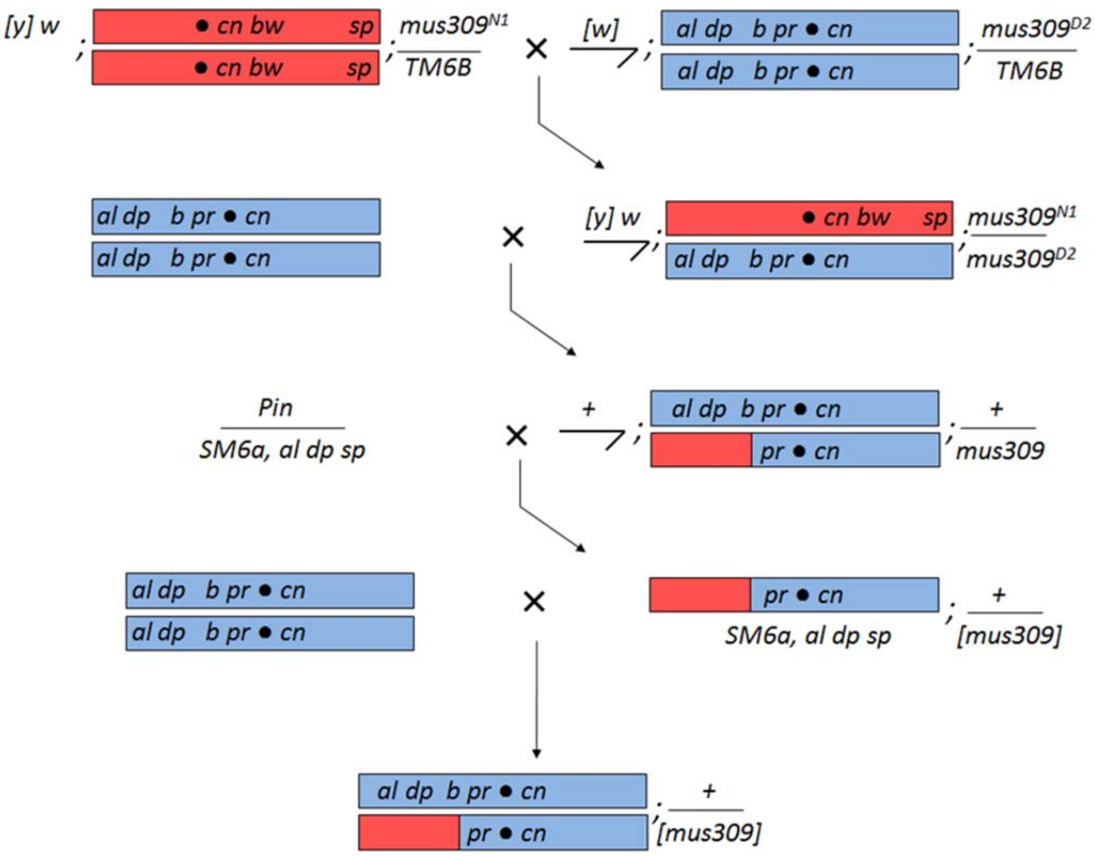


Figure 2.4. Cross scheme to visualize and recapitulate mitotic crossovers for future SNP sequencing. The initial steps are as described in Fig. 2.2, with the following relevant differences. Parental virgin females carry the Celera reference sequence chromosome, marked with *cn bw sp*. Males carry a slightly different marker chromosome, the alleles of which facilitate production of a crossover (CO) stock two generations later. Males also carry a different allele of *mus309*; the heteroallelic arrangement in the following generation prevents second-site mutations from affecting the experiment. COs are again generated in 2nd generation males, which are visualized by a cross to three to five marker chromosome virgin females (Cross 2). Again, one or both products of the CO can be recovered; one possible CO is shown here. If the fly in which the CO is visible is a male, the fly is crossed to a stock containing the *SM6a* 2nd chromosome balancer (Cross 3). The *al dp sp* markers carried on *SM6a* allows one to distinguish between the CO chromosome and the marker chromosome; in this way, balanced CO chromosomes are made into a stock. Flies from this stock can be crossed to marker chromosome flies to recapitulate the genotype of the original crossover fly; alternatively, flies from this stock can be crossed to each other to produce flies homozygous for the CO (Cross 4). Either type can be frozen in anticipation of future SNP mapping of the CO.

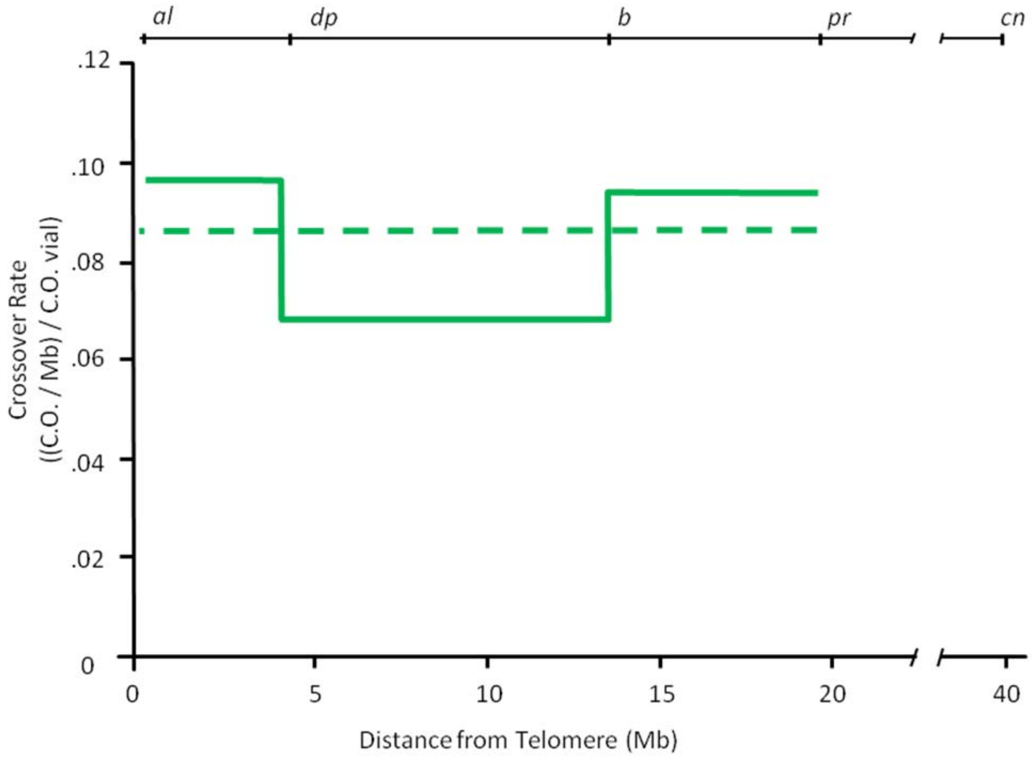


Figure 2.5. The rate of crossovers is non-uniform in a heteroallelic *mus309* background. The DmBLM crossover assay was repeated in a different genetic background, as depicted in Fig. 2.4. Crossovers were accrued between a marker chromosome carrying *al dp b pr cn*, and a reference chromosome with *cn bw sp*. The crossover rate displayed here only takes into account vials in which at least one crossover was detected. Crossover data was collected in vials of three to five females crossed to a single male; each vial therefore represents crossovers in the germline of one male. The solid green line represents males mutant for *mus309*^{N1}/*mus309*^{D2} (634 crossovers, 391 vials), and the dotted line is the average crossover rates over the entire 19.6 Mb interval.

The *mus309* mutant background provides information about the rate of endogenous breakage. While it would be useful to modify this assay – by adding aphidicolin – to obtain information about the rate of breakage due to an exogenous source, this would not be practical. Instead, I chose to use a genetic means of mimicking exogenous damage. I introduced replicative stress by using one mutant copy of *Polα*; 334 crossovers were analyzed from 157 *mus309*^{N1} *+/mus309*^{N1} *Polα* males. By doing so, I found that the rate of mitotic crossing over on *2L* was again significantly non-uniform ($P < 0.0001$). In most regions, the rate was elevated above the rate detected in the version of the assay that was designed to detect purely endogenous damage.

In both cases, the regions between *net* and *dp*, and between *b* and *pr*, appear to be more prone to DNA breakage than the other regions of *2L*.

I found that my assay often produced reciprocal crossovers. For example, if a crossover occurred between *b* and *pr*, the following generation might produce both *net* $dpp^{d\text{-}ho}$ *dp b* + + flies, as well as their + + + + *pr cn* siblings. Since this presented a unique opportunity to learn about DNA repair in a *mus309* mutant background by examining both products of a crossover, reciprocal crossover flies from the second iteration of the assay (the *al dp b pr cn* version) were crossed to produce multiple flies that recapitulated the original crossover (Fig 2.4). The DNA of seven of these flies has been submitted for high-throughput sequencing and SNP mapping.

Discussion

Non-uniform breakage indicates CFSs

I undertook this study to determine if *D. melanogaster* had CFSs, and to identify the location of such sites. To do so, I developed an assay that allowed me to locate sites of endogenous double-strand breaks (DSBs) by visualizing these sites as mitotic crossovers. In addition, I have laid a foundation for further studies that are expected to increase the resolution of this initial mapping by orders of magnitude.

I have demonstrated that the rate of mitotic crossing over, and therefore the rate of DNA damage repaired by a crossover, is significantly non-uniform on *2L*. This observation holds true in all backgrounds tested, and indicates that *2L* has regions that are more prone to DSBs than other regions. It is important that similar crossover distributions were observed in genetic backgrounds with different 2nd chromosomes, as it shows these damage-sensitive regions to not be an isolated phenomenon. These regions – specifically, *net-dp* and *b-pr* – therefore constitute CFSs. This is noteworthy, as this is the first report of CFSs in *Drosophila*. The apparent evolutionary conservation of CFSs as a feature of chromosomes is consistent with the report of regions analogous to induced CFSs in budding yeast (LEMOINE *et al.* 2005). Furthermore, the fact that I was able to locate regions prone to endogenous DSBs has important implications for our understanding of fragility.

Sources of breaks in the crossover assay

The natural and induced versions of the assay both detect DSBs associated with replication; the difference in the two methods lies in the initial source of the break. In the assay presented here, natural CFSs are regions that incur DSBs in cells that lack DmBLM. In general, endogenous DSBs are likely to be the result of single-strand DNA lesions, such as single-strand breaks, apurinic/apyrimidinic sites, and oxidation products (VILENCHIK and KNUDSON 2003). Such lesions can be converted to DSBs during DNA replication. Therefore, regardless of when the initiating lesion is acquired, the actual DSB is likely to be produced during S phase. In the context of *mus309* mutant cells, lesions that block replication fork progression are likely to be the main contributors to endogenous DSBs, because DmBLM likely acts to prevent the collapse of stalled forks (DAVIES *et al.* 2007). Such fork stalling can result from the presence of an abasic site on the leading strand (HIGUCHI *et al.* 2003). Without DmBLM to promote fork regression or stabilization of the fork, encountering such a lesion could lead to fork collapse and a DSB (MANKOURI and HICKSON 2007). Therefore, DmBLM mutant cells should not have any more stalled forks than wild type cells, but the forks that do stall should be much more likely to become DSBs. These DSBs, in turn, are repaired as mitotic crossovers, due to the role of DmBLM in preventing crossovers during DSB repair.

Breaks at induced CFSs, on the other hand, are the result of inhibition of the replicative polymerases. Such inhibition leads to increased ssDNA at the fork and an overall reduction in fork speed, and may result in polymerase-helicase uncoupling (ARLT and GLOVER 2010; LETESSIER *et al.* 2011). These features could lead directly to DSBs; for example, a nick on ssDNA would result in a one-ended DSB. Alternatively, these features, particularly polymerase-helicase uncoupling, could result in stalling of the replication fork. In cells with a reduced level of active polymerases but which are wild type in all other respects, such as cells treated with aphidicolin, the increased number of stalled forks may exceed the cell's ability to stabilize them all. This results in increased DSBs.

In cells with reduced Polα and mutant *mus309*, the crossovers observed are from breaks likely to be due to a combination of these effects. Polα reduction increases the number of forks that stall, and the lack of DmBLM results in a greater percentage of those stalled forks resulting in DSBs. The DSBs are then repaired to produce COs. The crux of the issue, of course, is where the forks stall. As previous studies have focused on induced CFSs, it was not yet known if the regions of fork stalling due to polymerase inhibition are similar to those in which stalling normally occurs. The similarity of the natural and induced damage distributions detected by my crossover assay implies that fork stalling occurs at the same sites regardless of the state of replication inhibition.

Implications of replication inhibition and endogenous breaks

The distribution obtained from flies mutant for both *mus309* and *Polα* is interesting, in that it is a similar shape as the distribution detected with wild type *Polα.* The regions from *net-dp* and *b-pr* appear to be fragile in both of these genetic backgrounds. The similarity of these distributions is consistent with the notion that regions affected by endogenous lesions – that is to say, the natural CFSs – are the same as those damaged while under replicative stress, the induced CFSs. This supports the hypothesis that DNA sequence plays a substantial role in CFS fragility. While the timing of replication, for example, is likely to be different in flies with reduced levels of Polα, the sequence of the chromosomes remained constant between the two backgrounds.

The implication that natural and induced CFSs may be the same is important, as it would indicate that CFSs are more than aphidicolin-sensitive regions. It may be, for example, that CFSs are not simply sensitive to the polymerase-slowing effects of aphidicolin alone; rather, it appears to be the combination of aphidicolin and a natural tendency for replication forks to stall in that region that leads to chromosome breaks. In this scenario, the aphidicolin serves to push the regions that are already sensitive to endogenous fork stalling above the threshold of detection. This is consistent with a study that found CFSs in human cells lacking ATR exhibited instability even in the absence of aphidicolin (CASPER *et al.* 2002).

This connects to previous work from our lab, in which the reduction of the levels of Polα was studied (LAROCQUE *et al.* 2007a). Mutation of *Polα* alone did not produce any detectable genome instability phenotypes. However, when *Polα* was mutant in a background mutant for *mei-41*, which encodes the *D. melanogaster* ortholog of ATR, increases in apoptosis, loss of heterozygosity, and male germline mitotic crossovers were detected. Mutation of *mei-41* by itself was also found to have a significantly higher rate of genome instability relative to wild type. This indicates another situation in which cells are prone to endogenous damage, and where this effect can be exacerbated by the replicative stress of Polα reduction.

Interestingly, in that same study, mutation of one of the mitotic cyclins, cyclin A (CycA), was found to rescue the apoptosis phenotype of *mei-41; Polα*/+ back to the levels observed in *mei-41* mutants. One interpretation of this result is that mutation of cyclin A slows the cell cycle, giving the cell more time to deal with the damage and stalled forks incurred from the reduction of *Polα* and lack of *mei-41*. *CycA* mutation could not, however, reduce the levels of apoptosis observed in *mei-41* mutants. This suggests that *Polα* and *mei-41*, both of which have been shown to induce fragile regions in other organisms even without aphidicolin treatment, may affect genome stability in different ways. Polα reduction appears to result in a deleterious effect that can be ameliorated with enough time; perhaps it causes replication forks to slow or stop in a manner that doesn't require MEI-41 to fix. On the other hand, flies mutant for *mei-41* do not appear to receive any benefit from a slower cell cycle. This effect of the *mei-41* mutation may be due to a possible DNA repair function that has been proposed for *mei-41*, beyond its role as a cell cycle regulator (LAROCQUE *et al.* 2007a; LAROCQUE *et al.* 2007b).

Connections between DmBLM & CFSs

The DmBLM mutation may have had additional effects beyond simply allowing us to visualize DSB repair events as mitotic crossovers. Human BLM has been shown to be associated with ultra-fine DNA bridges in normal cells (CHAN *et al.* 2007). The presence of these bridges is elevated in cells that lack BLM; it has been inferred that the bridges represent catenated, intertwined DNA between sister chromatids, and that BLM is involved in promoting the necessary decatenation. Strikingly, it was later shown that these ultra-fine bridges are associated with many CFSs after treatment with aphidicolin (CHAN *et al.* 2009). This supports a model in which CFSs are among the last portions of the genome to be replicated, and require BLM to disentangle from the sister chromatid. If replication is slowed by an exogenous agent, CFSs are left either entangled or unreplicated – consistent with work showing that BS cells have slow replication (RASSOOL *et al.* 2003). Either way, this could easily lead to DNA breaks. Indeed, spontaneous chromosome breaks in BS patients have been found to be significantly associated with CFSs (FUNDIA *et al.* 1995). In my assay, regions exceptionally prone to breaks in *mus309 Polα* mutants are the same as those prone to breaks in flies mutant for *mus309* alone. Due to the association of BLM with CFSs detected in human cells, the similarity between the distribution of induced and natural CFSs presented here suggests that this relationship is present in *Drosophila*, as well. As BLM has been proposed to prevent the collapse of stalled replication forks, this relationship provides further evidence that fork stalling occurs at CFSs even in the absence of replication inhibition.

It is interesting to note that the distribution of mitotic crossovers I detected is very different from the distribution of meiotic crossovers (*e.g.* MCVEY *et al.* 2007). For example, based on the meiotic crossover rate, the region distal to the centromere, *net-dpp*$^{d\text{-}ho}$, would be expected to have a very low rate; the region between *P-b* would be expected to have a high rate. I observed the opposite effect (Figs 2.2 and 2.5). This suggests that the relative elevation of mitotic crossovers in fragile regions is truly due to the repair of DSBs, and not due to any inherent propensity for forming crossovers.

Having determined the location of CFSs due to replication-associated damage, I aim to resolve these mitotic crossovers at a higher resolution. Examining the sites at a high resolution will allow me to determine if the sites I detected display clustering beyond what I was able to detect with my megabase-resolution phenotypic markers. Detection of such clustering would allow one to not only identify the boundaries of the fragile regions, but also to focus on the most fragile portion of the region in subsequent analyses of the causes of fragility. To facilitate such future studies, I have prepared crossovers for SNP mapping via high-throughput sequencing. This has entailed additional crosses of crossover flies, the creation of stocks containing balanced crossovers, and the freezing of flies in which I've recapitulated the genotype of the original crossover male (*e.g., al dp b pr cn*/CO) (Fig 2.4). I have commenced sequencing of reciprocal crossover products; as it is often difficult to recover both halves of a mitotic crossover in metazoans, this will give us a unique opportunity to study DNA repair in a *mus309* mutant background.

In summary, I have demonstrated that CFSs are present in *D. melanogaster*. Two fragile regions appear to be present on *2L*, based on natural and induced replication difficulties. Comparison of the results obtained with normal and reduced levels of Polα suggest that inhibition of replication pushes natural CFSs above the threshold of detection in tranditional CFS assays. Future studies will use higher resolution approaches to ascertain the distribution of endogenous DSBs within the *2L* fragile regions. I have laid the foundation for such studies by freezing CO flies in anticipation of SNP mapping of the COs via high-throughput sequencing. Such analyses will aid in determining the causes of fragility.

Materials and Methods

Drosophila stocks and genetics

Flies were reared on standard medium at 25° C, and virgined at 18° C overnight. The marker chromosome stock used for initial crossover experiments was *net* $dpp^{d\text{-}ho}$ *dp b pr cn;* $mus309^{N1}$*/TM6B*. Males of this stock were crossed to females of genotype *P{SUPor-P}GlcAT-*$S^{KG01446}$*;* $mus309^{N1}$*/TM6B*. Male progeny that were homozygous for $mus309^{N1}$ and heterozygous for the 2nd chromosome were crossed to *net* $dpp^{d\text{-}ho}$ *dp b pr cn* females; the progeny of that cross were scored for mitotic crossovers between the phenotypic makers. Crossovers that occurred between *dp* and *b* were further characterized via PCR to determine if they occurred proximal or distal to the *P* element. For crosses in which one copy of *Polα* was removed, the marker chromosome stock was changed to *net* $dpp^{d\text{-}ho}$ *dp b pr cn; ru* $mus309^{N1}$ *DNApol-α180 ca/TM6B*. I used the DEVIAT program developed by Mohamed Noor to perform bootstrapping to test if CO distributions were significantly non-uniform (CIRULLI *et al.* 2007).

Crossover flies designed to be used for SNP mapping were obtained in a similar fashion, with the following differences. Parental males were *al dp b pr cn/SM6a;* $mus309^{D2}$*/TM6B*, and parental females were *w; cn bw sp;* $mus309^{N1}$*/TM6B*. The second chromosome of the females was derived from the reference sequence stock, available from the Bloomington stock center. Male progeny that were heteroallelic for *mus309* and heterozygous for the 2nd chromosome were crossed to *al dp b pr cn* females; progeny of that cross were scored for mitotic crossovers.

If at least one crossover fly of that cross was a male, an attempt was made to make a balanced stock of the crossover chromosome. The crossover-bearing male was crossed to *y; Pin/SM6a, al dp sp*; the *al*, *dp*, and *sp* on *SM6a* were used in the following generation to distinguish the crossover chromosome from the marker chromosome. Siblings that carried both the crossover chromosome and *SM6a* were crossed to each other to make a stock.

If *al* and *dp* were both present on the initial crossover chromosome, there was a possibility that *sp* had been crossed off in an unrelated mitotic crossover. To avoid this situation, which would make it difficult to distinguish between the marker and crossover chromosomes, the male was first crossed to *net* $dpp^{d\text{-}ho}$ *dp* $wg^{Sp\text{-}1}$ *b pr cn/SM6a*. Male progeny of this cross that were not balanced on the 2nd chromosome were crossed to *y/y$^+$Y; Pin/SM6a, al dp sp* females, and the appropriate progeny were crossed to make a stock, as above.

Males that were to be used for SNP mapping via high-throughput sequencing were crossed to *al dp b pr cn*. Typically, this was done from a balanced crossover stock, but it could also be performed directly from the initial male that manifested the crossover. All progeny of the genotype *al dp b pr cn*/CO were collected and frozen at -80° C to await library preparation. The purpose of this approach was to generate multiple flies that had identical 2nd chromosome content, thus providing additional DNA for sequencing library preparation.

PCR analysis

The first crossover detection scheme used PCR to determine if crossovers between *dp* and *b* occurred proximal or distal to the *P* element. DNA was obtained via single-fly squishes (GLOOR *et al.* 1993). The primers were GTCTAGTGCCAGGCTACTCG and GCGGACCACCTTATGTTATTTC; the annealing step was 65° C, the extension step was 30 seconds, and 35 cycles were run.

Chapter III

Common Fragile Sites and Exogenous DNA Insertions

Introduction

I showed in Chapter II that *D. melanogaster* has common fragile sites (CFSs) on the left arm of chromosome *2*. This encouraged me to look for CFSs across the genome as a whole. I approached this study aiming to design a CFS detection assay that was both high-resolution and genome-wide. To accomplish this, I took advantage of a property that CFSs have in addition to – and likely because of – their propensity to incur DNA breaks.

It has been demonstrated that CFSs have a tendency to take up DNA from exogenous sources. This has been shown to be the case with aphidicolin-treated human/hamster hybrid cells, in which a selectable DNA cassette was found to preferentially integrate in FRA3B, a known human CFS (RASSOOL *et al.* 1991). In this study, the authors used fluorescence *in situ* hybridization to identify the chromosome band in which the integration took place. They found a significant hybridization in FRA3B in aphidicolin-treated cells; cells that had not been treated with aphidicolin had a more diffuse integration pattern, but did contain a site of non-random integration in the hamster portion of the DNA. It is unclear if the non-random integration in cells without aphidicolin was due to sequence characteristics of the construct, or if the integration site constitutes a natural CFS. Either way, the work represents the first experimental investigation of DNA integration at CFSs.

Similar tendencies to take up exogenous DNA, can be found in viral integrations (POPESCU *et al.* 1990). In this study, the authors reviewed the literature on integration sites of DNA-containing viruses, and found that the majority integrated in chromosome bands containing a CFS. Additional studies have found integrations of HPV16 within numerous CFSs (THORLAND *et al.* 2000; WILKE *et al.* 1996).

Non-random integrations at CFSs have also been detected in breast cancer cell culture (MATZNER *et al.* 2003). The MDA-MB-436 cell line was known to spontaneously express CFSs; the authors used fluorescent *in situ* hybridization to tag integrations of a selectable construct at these sites. Many of the integrations were found to co-localize to canonical CFSs, while the others integrated at other spots of known spontaneous breakage in the unstable cell line.

This propensity to take up exogenous DNA is likely due to the instability of CFSs, especially when under conditions that put the genome under stress; DSBs have been shown to take up ectopic DNA in *S. cerevisiae* (HAVIV-CHESNER *et al.* 2007; MOORE and HABER 1996). Several human CFSs were first able to have their sequence analyzed due to an approach based on the cloning of inserted DNA (MISHMAR *et al.* 1998; RASSOOL *et al.* 1996).

I designed an assay to take advantage of the integration-prone characteristic of fragility. My goal was to identify putative fragile regions, rather than to characterize known CFSs. I incorporated the idea of introducing a selectable construct to cells, but also included high-throughput sequencing (HTS) to efficiently identify integration sites at a high resolution. The assay can be used with or without aphidicolin, and may thus be used to gauge the effect of replicative stress on DNA integration.

My assay has been successfully used to identify multiple DNA integration sites, thus providing evidence for the possible location of CFSs. Analysis of these sites has provided information about the type of DNA repair used to integrate the foreign DNA, as well as information about the factors that contribute to fragility.

Results

The integration assay identifies putative fragile regions

I designed a novel assay to locate integration sites of a linear, selectable DNA construct into the genome of *D. melanogaster* S2 cells (Fig 3.1). Briefly, the construct was transfected into cells in the presence of aphidicolin, and the construct was selected for; the insert-containing genomic DNA was then harvested and used for either HTS or TOPO cloning; either method could be used to identify the location of multiple insertions with a very high resolution. The HTS method relies on paired-end sequencing to identify fragments that contain both insert and genomic DNA, while TOPO cloning simply sequences across the junction.

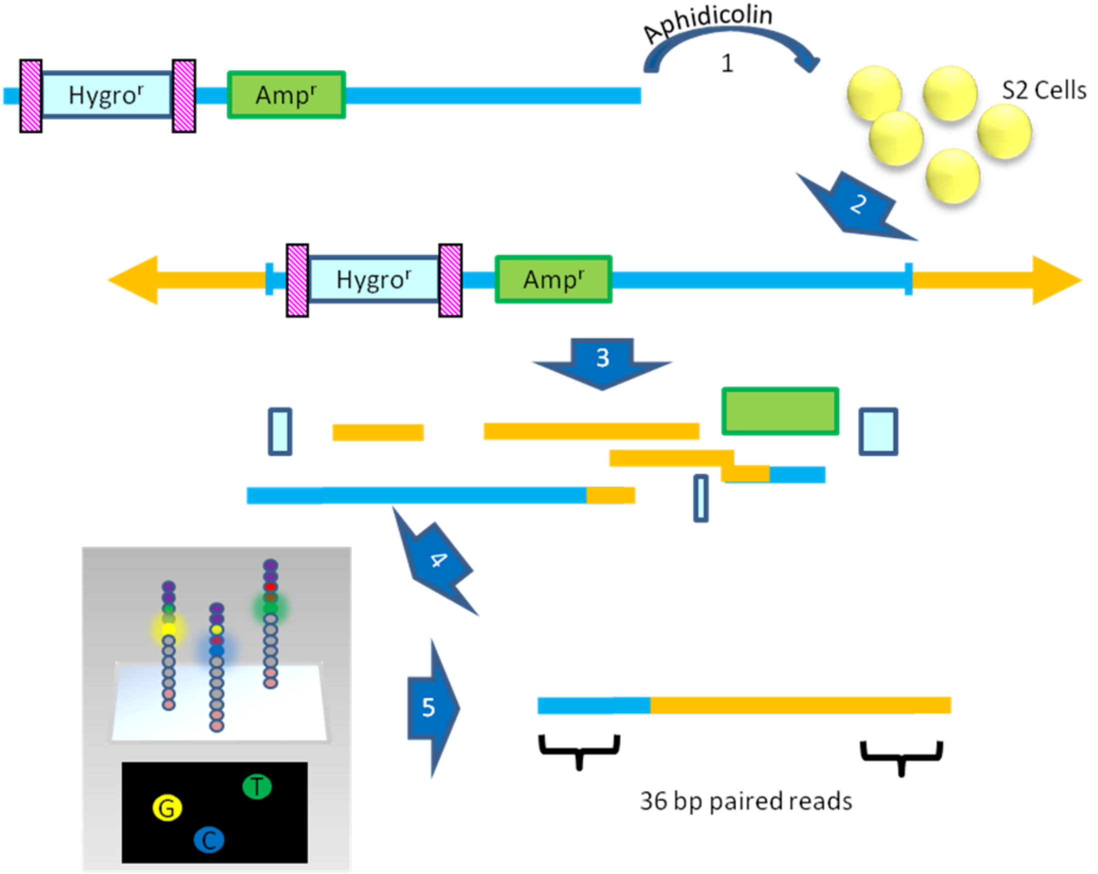


Figure 3.1. Cell-based DNA integration detection scheme. I have designed an assay that will allow one to map DNA integration sites across the entire genome, indicating putative fragile regions. A modified, linear version of the pCoHygro vector was used. It contains a gene conferring hygromycin resistance (light blue box), flanked by Drosophila *gypsy* insulators (pink boxes). The construct is transfected into S2 cells, with or without aphidicolin treatment (1). Stable integrations (2) can be selected by treating cells with hygromycin, or cells can be harvested as soon as they regain confluence. Genomic DNA is depicted here as orange lines. The DNA is extracted from the cells and sheared to ~400 bp fragments via sonication (3). Sequencing adapters are added to the ends of the fragments, allowing them to be sequenced on the Genome Analyzer IIx (Illumina; 4). The resulting 36 bp paired-end reads allow one to pinpoint the integration site of multiple insertions, indicating potential CFSs on a genome-wide scale (5).

When cells transfected with 500 ng of construct were analyzed via HTS, I was able to detect 23 unique, unambiguous integrations of the construct with an initial resolution of less than 400 bp (Fig 3.2 A, Table 3.1). The high resolution is due to the size to which DNA fragments are sheared during HTS library prep. Five additional insertion events from this initial pool were identified, but due to their integration into natural transposable elements, their exact location could not be ascertained from HTS alone. The possible sites of their integration are presented, relative to the *D. melanogaster* reference sequence (Fig 3.2 B).

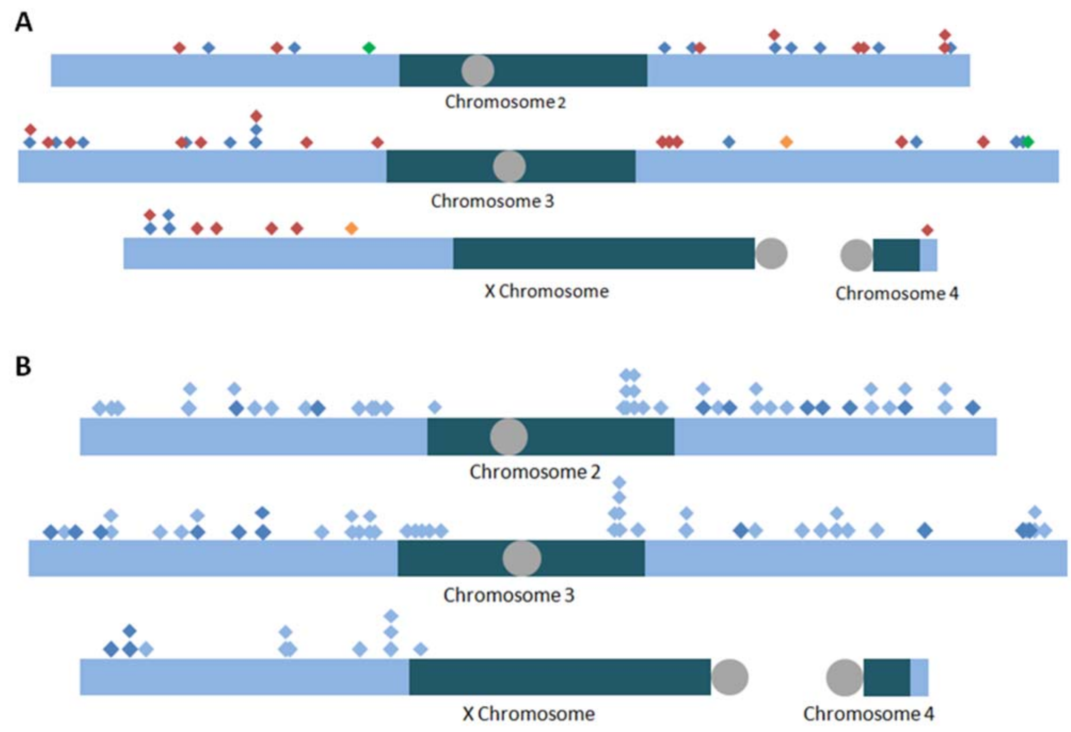


Figure 3.2. Integration of aphidicolin-induced DNA inserts. Integration events of pCoHygro into S2 cell DNA in the presence of aphidicolin were mapped with a variety of techniques. Those that could be unambiguously mapped to a single location are presented here relative to the *D. melanogaster* reference sequence assembly (A). Depicted are insertions detected in cells transfected with 500 ng of linear pCoHygro and selected with hygromycin for 19 days, detected with GA IIx sequencing (23 inserts, blue diamonds) or splinkerette PCR and TOPO-TA cloning (2 inserts, green), and cells transfected with 5 ng of linear pCoHygro and grown for 7 days without selection, detected with GA IIx sequencing of splinkerette PCR products (27 inserts, red), or TOPO-TA cloning (2 inserts, orange). The possible integration shites of the five ambiguous integrations from the 500 ng GA IIx sample are presented in (B). Every possible integration site of those five inserts are depicted as light blue diamonds; unambiguous inserts from this pool are blue diamonds, as in (A). Stacked diamonds are inserts within 20-155 kb of each other. Light blue blocks, euchromatin; dark blue blocks, heterochromatin; circles, centromeres.

Name	Arm	Start	End
032509ins_1	2L	10358756	10359131
032509ins_2	2L	16288849	16289224
032509ins_3	2R	15531682	15532057
032509ins_4	2R	20494516	20494891
032509ins_5	2R	2644710	2645085
032509ins_6	2R	8358685	8359060
032509ins_7	2R	723361	723736
032509ins_8	2R	11474467	11474842
032509ins_9	2R	9474495	9474870
032509ins_10	3L	4152476	4152851
032509ins_11	3L	2301634	2302009
032509ins_12	3L	15996392	15996767
032509ins_13	3L	16026055	16026430
032509ins_14	3L	14253214	14253589
032509ins_15	3L	11227946	11228321
032509ins_16	3L	490738	491113
032509ins_17	3R	5563623	5563998
032509ins_18	3R	26717239	26717614
032509ins_19	3R	19040195	
032509ins_20	3R	26235894	26236269
032509ins_21	X	1093354	1093729
032509ins_22	X	2475619	2475994
032509ins_23	X	2607150	
032509ins_24	3R*	27093353	27093358
032509ins_25	2L	Histone locus	

Table 3.1. Location of unambiguous integrations in aphidicolin-treated cells selected with hygromycin. These integrations came from the first electroporation of S2 cells with 500 ng of the construct. Displayed are the possible ranges that contain the insert-genome junction for a given integration. In instances in which the precise location was determined, only one number is given. An asterisk indicates that a precise junction of the insert and genome was determined, but that there was microhomology. The colors correspond to those used in Fig. 3.2: blue inserts were detected *via* HTS, and the green inserts were detected *via* TOPO cloning of splinkerette PCR products. The green insert on *2L* landed in the histone locus, but could not be placed to an unambiguous location. All locations refer to release 5 of the *D. melanogaster* genome.

I found that, even when transfecting 100-fold less DNA, I was able to detect unambiguous integrations. In this iteration of the experiment, splinkerette PCR was used to enrich for insert-containing sequences after the extracted DNA had been sheared (Fig 3.3). Splinkertte PCR involves the annealing of a "splink" adapter to both ends of each sheared fragment (DEVON *et al.* 1995; UREN *et al.* 2009). The hairpin structure on the adapter is designed to prevent “end-repair” priming during PCR, in which unligated DNA anneal and amplify fragments that do not contain the insert. In the first round of PCR, one of the two primers used matches the sequence in the single-stranded region of the splink; that is, it has no complementary sequence, and can therefore not initiate DNA synthesis. The other primer used is an insert-specific primer, which accomplishes first-strand synthesis only in fragments that contain the corresponding portion of the insert. Synthesis from the insert-specific primer produces the complementary sequence that the splink primer needs to continue the PCR. Therefore, even though the majority of the DNA fragments are nothing but genomic DNA, splinkerette PCR can enrich for DNA fragments that contained insert DNA. In this manner, I detected an additional 27 unambiguous integrations in cells transfected with 5 ng of construct (Fig 3.2 A, Table 3.2).

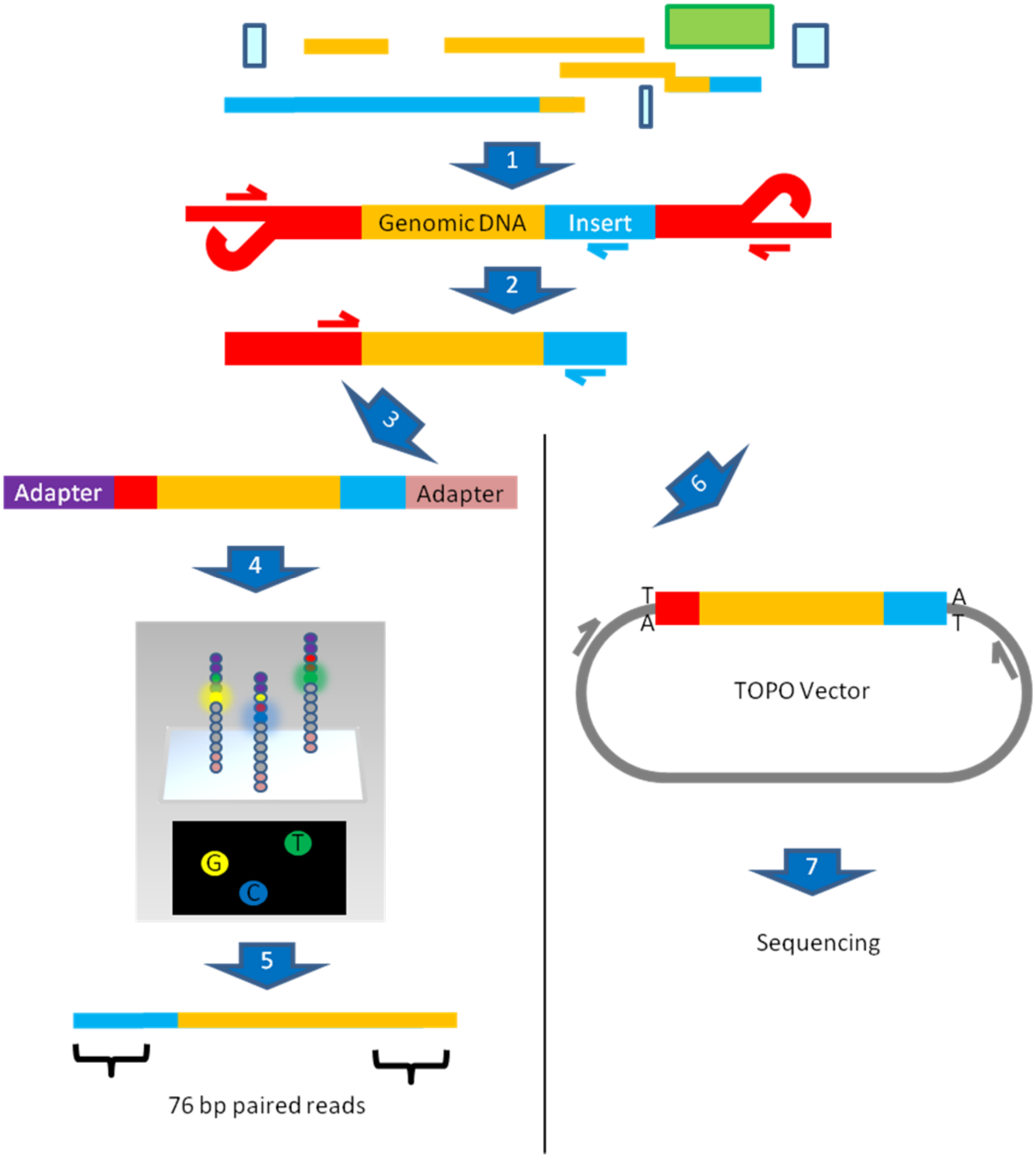


Figure 3.3. Splinkerette PCR enriches for insert-containing fragments. Following sonication of the DNA, blunt splinkerette adapters (red) are ligated to the ends of the sheared fragments. An insert-specific PCR primer is employed to enrich for insert-containing fragments, while the hairpin on the splinkerette adapters prevent end-repair priming (1). A second round of PCR further amplifies sequences of interest, while shortening the terminal non-genomic DNA (2). Sequencing adapters are added to the ends of the PCR products, allowing them to be sequenced on the GA IIx (3,4). The resulting 76 bp paired-end reads allow us to pinpoint the integration site of multiple insertions (5). Alternatively, the product of the nested PCR from (2) can be cloned into a TOPO-TA vector and transformed into *E. coli* (6). Colony PCR is used to detect successful cloning, and sequencing of the PCR products facilitates detection of the insert-genome junction (7).

Name	Arm	Start	End
100109ins_1	2L	8348443	8348693
100109ins_2	2L	15047152	15047402
100109ins_3	2R	3194050	3194300
100109ins_4	2R	8514521	8514771
100109ins_5	2R	14124686	14124936
100109ins_6	2R	14516313	14516563
100109ins_7	2R	19989569	19989819
100109ins_8	2R	20109041	20109291
100109ins_9	3L	619982	620232
100109ins_10	3L	1731522	1731772
100109ins_11	3L	3245748	3245998
100109ins_12	3L	10905065	10905315
100109ins_13	3L	12224632	12224882
100109ins_14	3L	15976870	15977120
100109ins_15	3L	19503104	19503354
100109ins_16	3L	24379771	24380021
100109ins_17	3R	758293	758543
100109ins_18	3R	1252771	1253021
100109ins_19	3R	1822864	1823114
100109ins_20	3R	17990129	17990379
100109ins_21	3R	23872548	23872798
100109ins_22	4	781501	781751
100109ins_23	X	1237783	1238033
100109ins_24	X	4522381	4522631
100109ins_25	X	5956328	5956578
100109ins_26	X	10021582	10021832
100109ins_27	X	11881239	11881489
100109ins_28	3R	9736291	
100109ins_29	X	15867840	

Table 3.2. Location of unambiguous integrations in aphidicolin-treated cells enriched *via* splinkerette PCR. These integrations came from the second electroporation of S2 cells, which used 5 ng of the construct. Displayed are the possible ranges that contain the insert-genome junction for a given integration. In instances in which the precise location was determined, only one number is given. The colors correspond to those used in Fig. 3.2: red inserts were detected *via* HTS of splinkerette PCR products, and the orange inserts were detected *via* TOPO cloning of splinkerette PCR products. All locations refer to release 5 of the *D. melanogaster* genome.

The products of splinkerette PCR were also used for TOPO cloning, a system that allows direct cloning of PCR products. By doing so, I was able to maximize the resolution of the insert-genome junction, at the cost of some of the throughput. TOPO cloning of the 500 ng transfection revealed two additional inserts: an unambiguous integration on 3R, and an insertion in the repetitive DNA of the histone locus on *2L* (Fig 3.2 A, Table 3.1). The 5 ng transfection sample was found to contain two additional unambiguous integrations (Fig. 3.2 A, Table 3.2).

The assay was also used to detect integrations in the absence of aphidicolin. The procedure was similar to that described above, with three exceptions. First, the inserted construct was slightly different: I used a version that did not contain the gypsy insulators. As these regions are *D. melanogaster* sequences, I removed them to eliminate potential integration bias, and to simplify insertion mapping. Second, the S2 cells were not exposed to aphidicolin. Third, in order to maximize the heterogeneity of the final pool of inserts, electroporated cells were split into twelve wells of 24-well plates immediately after transfection. By performing hygromycin selection in twelve wells instead of one, the total cell population could not become completely homogenous due to selective growth. Therefore, the diversity of inserts at the end of the selection process was anticipated to be greater. The high-throughput sequencing analysis of inserts generated in such a manner allowed me to detect seven integrations that took place without the presence of aphidicolin (Fig. 3.4, Table 3.3).

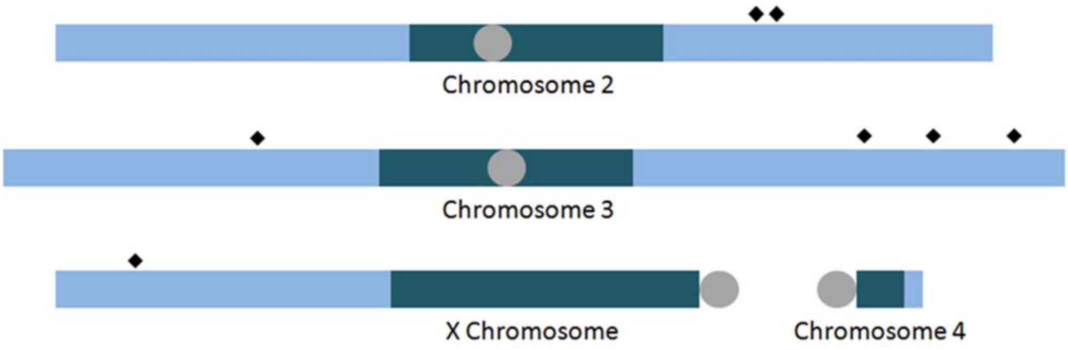


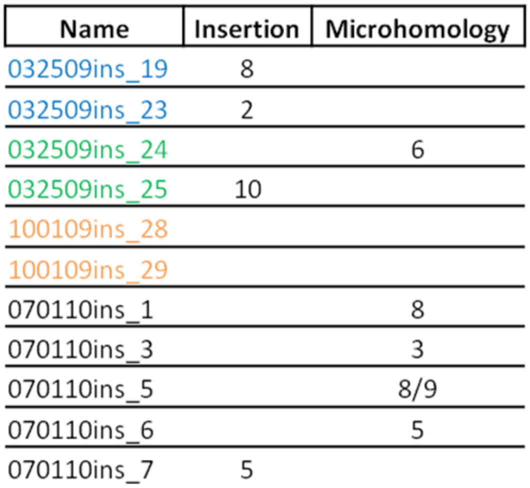
Figure 3.4. Integration of DNA inserts without aphidicolin. Integration events of pCoHygro into S2 cell DNA were mapped with HTS; those that could be unambiguously mapped to a single location are presented. Depicted are insertions detected in cells transfected with 500 ng of linear pCoHygro and selected with hygromycin for 15 days, detected with GA IIx sequencing. Light blue blocks, euchromatin; dark blue blocks, heterochromatin; circles, centromeres.

Name	Arm	Start	End
070110ins_1	2R*	5457756	5457763
070110ins_2	2R	6847841	6848041
070110ins_3	3L*	16896695	16896697
070110ins_4	3R	15049148	15049348
070110ins_5	3R*	19894015	19894023
070110ins_6	3R*	25573134	25573138
070110ins_7	X	4546351	

Table 3.3. Location of unambiguous integrations in cells selected with hygromycin. These integrations came from the third electroporation of S2 cells, which used 500 ng of the construct but no aphidicolin. Displayed are the possible ranges that contain the insert-genome junction for a given integration. In instances in which the precise location was determined, only one number is given. An asterisk indicates that a precise junction of the insert and genome was determined, but that there was microhomology. The color corresponds to that used in Fig. 3.5: all inserts were detected *via* HTS. All locations refer to release 5 of the *D. melanogaster* genome.

Integration analysis: insert-genome junctions

In analyzing the integrations, I first focused on the sequence of insert-genome boundaries. The uptake of exogenous DNA requires some means of DNA repair, and the sequence of the junctions provides a “signature” to suggest what type of repair was used. I took two approaches to analyzing the junctions. The first was to use sequence data obtained from high-throughput sequencing to design primers to amplify insertion junctions that were known at the 400 bp resolution. The PCR products were then sequenced; two integrations were analyzed in this way (Table 3.4). The second approach was to examine the initial sequencing data to see if one of the sequencing reads spanned the junction. The four integrations that had been detected by TOPO cloning necessarily had one junction that was known to single-bp resolution, and five of the seven non-aphidicolin inserts had at least one read that spanned the junction (Table 3.4).



Name	Insertion	Microhomology
032509ins_19	8	
032509ins_23	2	
032509ins_24		6
032509ins_25	10	
100109ins_28		
100109ins_29		
070110ins_1		8
070110ins_3		3
070110ins_5		8/9
070110ins_6		5
070110ins_7	5	

Table 3.4. Sizes of insertions and microhomology at integration-genome junctions. All 11 junctions analyzed are listed. The size of these events is given in base pairs. The color corresponds to those used in Figs. 3.2 and 3.5; colored integrations came from cells treated with aphidicolin, and black integrations came from cells that were not. The "8/9" indicates imperfect microhomology, in which only 8 of 9 possible bases between the construct and the reference genome matched.

Analysis of the junctions revealed differences between aphidicolin and non-aphidicolin integrations. Non-aphidicolin inserts had a more frequent use of microhomology, in which the junction contains a short (1-10 bp) stretch of nucleotides that match both the insert and the genomic sequence (Fig 3.5, Table 3.4)(MCVEY and LEE 2008). It is unclear if the single instance of imperfect microhomology is due to a mismatch or a SNP in the S2 DNA relative to the reference sequence. Aphidicolin integrations had a higher rate of small (1-10 bp) insertions between the construct and genomic DNA, the source of which is unclear. These data are consistent with an interpretation in which, in the absence of aphidicolin, microhomologies play a substantial role in determining the precise location of integration; when aphidicolin is present, the criteria are not as strict.

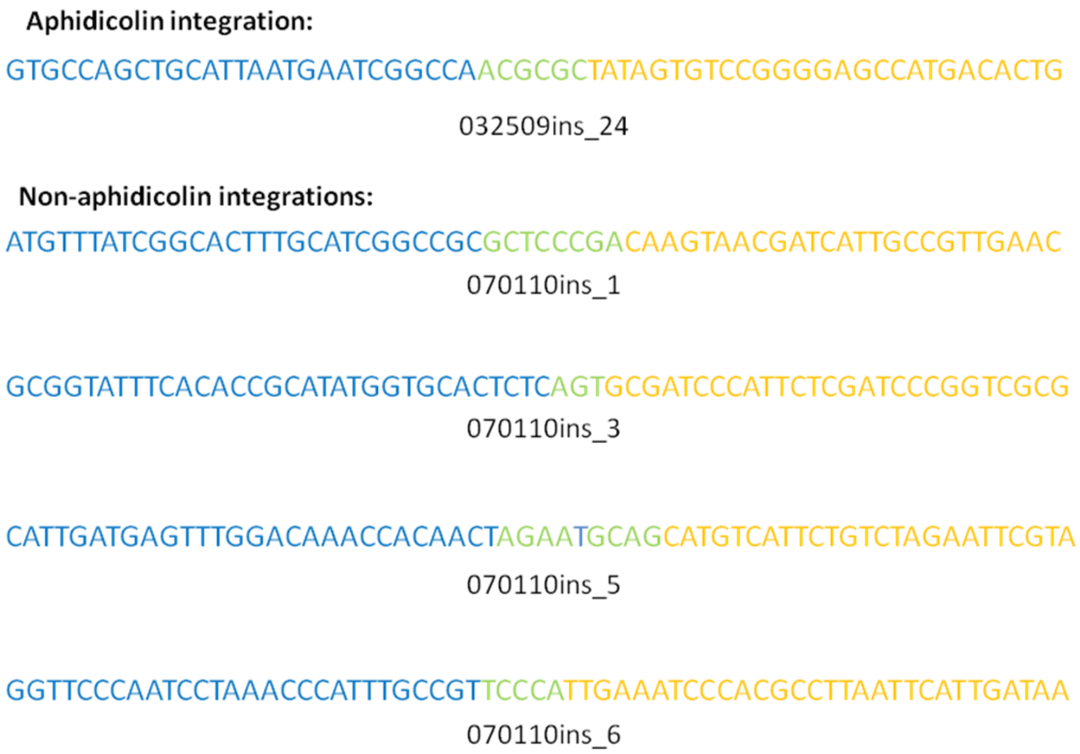

Aphidicolin integration:
GTGCCAGCTGCATTAATGAATCGGCCAACGCGCTATAGTGTCCGGGGAGCCATGACACTG
032509ins_24
Non-aphidicolin integrations:
ATGTTTATCGGCACTTTGCATCGGCCGCGCTCCCGACAAGTAACGATCATTGCCGTTGAAC
070110ins_1
GCGGTATTTCACACCGCATATGGTGCACTCTCAGTGCGATCCCATTCTCGATCCCGGTCGCG
070110ins_3
CATTGATGAGTTTGGACAAACCACAACTAGAATGCAGCATGTCATTCTGTCTAGAATTCGTA
070110ins_5
GGTTCCCAATCCTAAACCCATTTGCCGTTCCCATTGAAATCCCACGCCTTAATTCATTGATAA
070110ins_6

Figure 3.5. Microhomology at insert-genome junctions. The sequence of the insert-genome junctions that were detected to have microhomology is presented above. Only 070110ins_5 had imperfect microhomology, represented by the T that corresponds to the insert, but not the reference genome. Insert DNA is represented in blue, genomic DNA in orange, and microhomology in green.

In situations where I did not determine the exact junction, I used the size of the HTS library band, adapter length, and read length to determine the minimal range that could contain the junction. For example, the band size of the 500 ng aphidicolin-treated library was about 500 bp, the read length was 36 bp from either end, and both sides of each fragment had 33 bp worth of Illumina adapter attached. The amount of unknown DNA in the 500 bp span was therefore:

$$500-(2*36)-(2*33) = 362 \text{ bp}$$

For a conservative estimate, I rounded the possible junction range up to 375 bp (Table 3.1). Using similar calculations, I determined the range of the 5 ng sample to be 250 bp (Table 3.2). In the two cases in which the junctions of the 500 ng non-aphidicolin sample were not already known, the possible range was 200 bp (Table 3.3).

In at least two of the three transfections presented here, the position of the insert end of the paired-end reads indicates that the integration did not involve the entire plasmid (Fig 3.6). The inserts recovered from the splinkerette-enriched sample, shown in red, are likely from intact constructs – the reads could all theoretically fit the end of the construct within the space of the 400 bp fragment used for sequencing (Fig. 3.6 B). The other two samples, however, have construct reads that are kilobases away from either end of the construct, indicating that only part of the construct was integrated at that site (Fig 3.6 A, C). This could be indicative of fragmentation or degradation of the construct prior to integration, or the use of homologous recombination to take up only part of the linear construct.

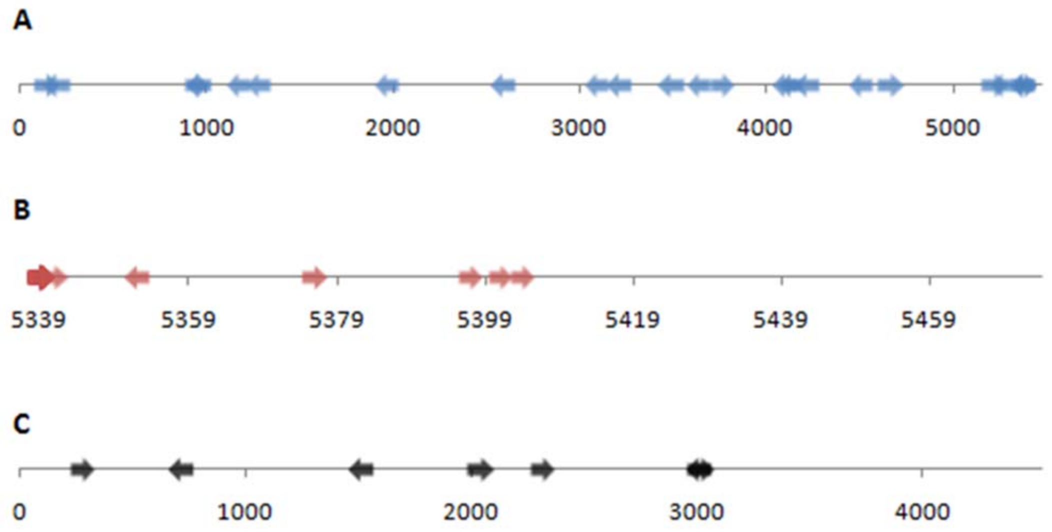


Figure 3.6. Position of insert reads suggest fragmented integrations. The position and orientation of the unambiguous sequencing reads are depicted. Blue arrows are reads from the 500 ng sample + aphidicolin (A). Red arrows are the reads from the 5 ng sample + aphidicolin, enriched with splinkerette PCR. The PCR selects for the terminal 139 bases, depicted here (B). Black arrows represent reads from the 500 ng sample that was not treated with aphidicolin. This version of the construct did not contain the *gypsy* adapters; hence, it is shorter.

Integration analysis: local genomic environment

As noted above, the location of integration events is expected to correlate with *D. melanogaster* CFSs. Therefore, the most interesting regions from this experiment, with regard to CFSs, are those in which multiple insertion events cluster. However, every integration site is informative in the sense that there was something about that region that was amenable to taking up exogenous DNA. I analyzed these sites to determine what features gave the sites this fragile characteristic. I aimed to take advantage of my high-resolution, high-coverage approach to CFS characterization by analyzing the local genomic environment of the integration sites. This was accomplished largely by using data from the modENCODE project and leveraging the Galaxy bioinformatics framework (BLANKENBERG *et al.* 2010; CELNIKER *et al.* 2009; EATON *et al.* 2011; GOECKS *et al.* 2010).

The inserts were recovered at a relatively high rate per experiment, with 53 inserts coming from only two transfections involving aphidicolin, and an additional seven insertions from a transfection without aphidicolin. However, because we have little *a priori* indication of where *D. melanogaster* CFSs might be located, analysis of what makes these regions fragile would benefit from additional integration events. Future iterations of the integration assay will no doubt assist in this aim. Still, integration in the presence of aphidicolin does constitute evidence of chromosome fragility. With that in mind, I analyzed my integrations in terms of genomic factors that have been proposed to contribute to fragility. This was typically done by comparing the characteristics of my two integration sets – those treated with and without aphidicolin – to corresponding datasets of identically-sized intervals placed at locations determined by a random number generator. The random regions were designed to take the copy number variation of S2 cells into account (ZHANG *et al.* 2010).

Name	Timing	Distance	Density	% GC	Name	Timing	Distance	Density	% GC
032509ins_1	0.7981	5150	13	0.38	100109ins_1	1.5820	13581	12	0.54
032509ins_2	0.2517	443	16	0.52	100109ins_2	0.1539	7540	7	0.45
032509ins_3	0.2518	25119	2	0.41	100109ins_3	-0.1581	5530	5	0.5
032509ins_4	0.1336	6920	9	0.43	100109ins_4	0.7590	4871	12	0.52
032509ins_5	-0.1446	0	9	0.48	100109ins_5	-0.1049	52875	9	0.53
032509ins_6	0.4473	17670	2	0.54	100109ins_6	0.7842	9518	11	0.47
032509ins_7	0.1799	6520	3	0.5	100109ins_7	-0.1600	11882	13	0.44
032509ins_8	-0.4958	15381	3	0.58	100109ins_8	-0.0804	33749	9	0.4
032509ins_9	0.2581	5307	6	0.51	100109ins_9	0.2897	2728	21	0.43
032509ins_10	0.2330	13017	10	0.58	100109ins_10	-0.1006	266	5	0.46
032509ins_11	-0.0575	35559	3	0.42	100109ins_11	1.3318	252	19	0.45
032509ins_12	0.1651	761	12	0.44	100109ins_12	0.1136	12441	12	0.42
032509ins_13	0.6378	3428	14	0.4	100109ins_13	-0.7479	34183	3	0.4
032509ins_14	-0.1330	13558	2	0.46	100109ins_14	-0.0505	4304	10	0.42
032509ins_15	1.4822	10963	8	0.45	100109ins_15	0.4495	26658	3	0.37
032509ins_16	-0.6884	14656	10	0.38	100109ins_16	-0.3902	149037	0	0.35
032509ins_17	0.1433	16281	9	0.42	100109ins_17	-0.4151	18944	8	0.37
032509ins_18	0.1489	2149	9	0.47	100109ins_18	-0.5541	39726	7	0.42
032509ins_19	0.1998	13095	13	0.49	100109ins_19	-0.5959	9956	2	0.4
032509ins_20	1.2705	21214	5	0.44	100109ins_20	-0.4261	141157	0	0.42
032509ins_21	0.2630	147577	0	0.48	100109ins_21	-0.7628	38993	5	0.48
032509ins_22	0.9244	6397	16	0.44	100109ins_22	-0.5389	63051	5	0.36
032509ins_23	0.9874	698	17	0.55	100109ins_23	0.1683	55636	1	0.41
032509ins_24	0.1463	16613	10	0.62	100109ins_24	0.9527	37871	8	0.41
					100109ins_25	0.5187	24712	5	0.48
					100109ins_26	-0.7984	40882	3	0.44
					100109ins_27	0.3733	24734	4	0.59
					100109ins_28	-0.7664	8724	6	0.26
					100109ins_29	1.5389	39916	5	0.46

Table 3.5. Local genomic environment of aphidicolin-induced inserts. All values were determined using tracks from the modENCODE genome browser. Timing is the timing of replication from the S2 replication timing track. Values are the $\log_2$ ratio of early to late replicating sequences at the probe nearest the insertion window; positive values indicate early replication. Distance is the distance, in base pairs, to the nearest origin of replication, determined from the S2 dOrc2 track. A 0 indicates that the window overlapped an origin. Density is the number of origins within 100 kb of either side of the insertion window, again determined by S2 dOrc2 signals. %GC is the GC content of the insertion window, calculated using the EMBOSS geecee tool in Galaxy.

Name	Timing	Distance	Density	% GC
070110ins_1	0.6965	4661	7	0.43
070110ins_2	0.0691	61034	2	0.41
070110ins_3	-0.6700	12198	9	0.56
070110ins_4	0.1697	0	7	0.47
070110ins_5	-0.2761	20793	8	0.43
070110ins_6	0.7255	1033	28	0.4
070110ins_7	0.5577	14710	7	0.58

Table 3.6. Local genomic environment of inserts incurred without aphidicolin. All values were determined using tracks from the modENCODE genome browser. Timing is the timing of replication from the S2 replication timing track. Values are the $\log_2$ ratio of early to late replicating sequences at the probe nearest the insertion window; positive values indicate early replication. Distance is the distance, in base pairs, to the nearest origin of replication, determined from the S2 dOrc2 track. A 0 indicates that the window overlapped an origin. Density is the number of origins within 100 kb of either side of the insertion window, again determined by S2 dOrc2 signals. %GC is the GC content of the insertion window, calculated using the EMBOSS geecee tool in Galaxy.

Sequence composition: In humans, CFSs tend to be AT-rich, although no specific motifs have been found to be associated with fragility (LUKUSA and FRYNS 2008). The 100 bp flanking regions surrounding my inserts tended to consist mostly of AT content. The means of the insertion windows were significantly different than the means of the random sites for aphidicolin-treated cells, but not for non-aphidicolin cells (aphidicolin-treated: $P = 0.0145$; non-aphidicolin: $P = 0.3060$, Mann-Whitney U test). On average, the aphidicolin-treated cells had a higher GC content than the random regions (45% *vs*. 42%).

Replication timing: CFSs in other models tend to be late-replicating, and may be origin-poor. It has been reasoned that these regions are fragile because they are unable to replicate their DNA by the end of S phase, especially in the presence of replication-inhibiting aphidicolin.

Taken as a whole, I found considerable variability in the timing of replication of my integration sites (Tables 3.5, 3.6). The mean replication timing did not appear to differ from that of the random regions (aphidicolin-treated: $P = 0.0623$; non-aphidicolin: $P = 0.4924$, t-test). I also calculated the distance to the nearest origin of replication, which I identified based on the peak calls of the S2_dOrc2 subtrack of the modENCODE ChIP-Seq dORC2 track. This track was generated by immunoprecipitation of the origin recognition complex subunit 2, which binds to origins of replication. This distance varied substantially between different inserts, from those that overlapped an origin to those that were almost 150 kb away. Those that were furthest away from origins did not necessarily have the latest replication, however, speaking to the complexity of the replication protocol of the cell (Tables 3.5, 3.6). The mean distance of the insertion windows was not significantly different from that of the random regions (aphidicolin-treated: $P = 0.1665$; non-aphidicolin: $P = 0.3176$, Mann-Whitney U test).

The local density of origins of replication was also considered. I counted the number of origins, represented by dOrc2 peaks, that were within 100 kb of either side of the initial search window. I did not detect any significant difference between means of the aphidicolin-treated and random samples (aphidicolin-treated: $P = 0.4126$; non-aphidicolin: $P = 0.9491$, t-test).

Discussion

I have used a high-resolution, genome-wide DNA integration assay in *D. melanogaster* cells to identify regions exceptionally prone to taking up exogenous DNA, a characteristic of CFSs. In doing so, I detected evidence that suggests S2 cells may use different repair mechanisms to integrate exogenous DNA, depending on whether or not they are under replicative stress. My analysis of integration sites suggests that the causes of fragility are likely due to a combination of factors, as no single factor I investigated was sufficient to explain the fragility of the insertion sites.

Aneuploidy in S2 cells

S2 cells have a stable genotype, but are aneuploid for at least 43 Mb of the genome and contain many rearrangements (ZHANG *et al.* 2010). Because of this, there is not a sequence assembly for S2 cells. This makes analyzing the relative location of inserts difficult – that is, unknown rearrangements prevent me from being able to determine the extent of clustering. However, this has little effect on the validity of individual inserts, due to the high resolution to which the sites have been mapped. That is, the features near each integration can be reliably determined, but the distance between integration events cannot.

A similar issue affects the pool of ambiguous inserts, all of which occurred in natural transposable elements (Fig. 3.2 B). The location of transposable elements in S2 cells may differ from those of the reference sequence, so the potential insertion sites depicted in Figure 3.2 may not be accurate. What Figure 3.2 B serves to illustrate, however, is that the nature of transposable elements means that the five ambiguous inserts have numerous possible integration sites.

Replication inhibition may influence DNA repair

The insert-genome junctions analyzed here suggest that replicative stress affects the relative usage of microhomologies in the integration of exogenous DNA. Integrations in cells that had been treated with aphidicolin appeared to use microhomology less frequently than cells not treated with aphidicolin. In the absence of aphidicolin, exogenous DNA may be easiest to integrate in genomic regions with similar sequence. By adding aphidicolin, there may be a relaxing of the requirement for similar sequence, and/or an increase in the number of potential integration sites. This may occur because of additional replicative stress introduced by aphidicolin, in which the inhibition of replicative polymerases leads to chromosomal gaps and breaks that could take up exogenous DNA during repair. The results here suggest that this is an issue worthy of additional study. Future experiments will examine additional insert-genome junctions, including those that exist on the side of the integration not detected by HTS.

Integration data suggests fragmentation of the construct

The location of sequencing reads within the construct suggests that many integration events did not involve the entire construct (Fig. 3.6). In cases in which the hygromycin resistance gene was not integrated, it is likely that another integration that did contain hygromycin resistance occurred in the same cell, allowing the cell to survive selection. The integration of only part of the construct is consistent with fragmentation or partial degradation of the construct prior to integration, or the use of homologous repair to take up only part of the construct. Information from the insert junctions suggest that the samples treated with aphidicolin appear to have an infrequent use of microhomology at the insert-genome junction (Table 3.4). In this instance, fragmentation of the insert is a more plausible explanation than homologous repair. For the black, non-aphidicolin sample, where microhomology is more prevalent, the distinction is not as clear. However, as the sections of homology appear to be relatively small, repair involving fragmented DNA is likely a more plausible explanation than homologous repair.

Interpretation of insert environment

The possibility of multiple factors affecting integration underscores the complexity of chromosome fragility. Instability at mammalian CFSs is likely due to a combination of several factors, and the results of my analysis of the S2 cell integration sites are consistent with a model in which *Drosophila* CFSs are just as complex. From the analyses I have carried out so far, there does not seem to be a specific, single factor that these integration sites have in common, and which explains why they appear to be fragile. It is interesting to note, however, that the insert regions of aphidicolin-treated cells have a significantly higher GC content than random regions of equal size ($P = 0.0145$). The average GC content of said regions is 45%, consistent with the low GC content of human CFSs (SCHWARTZ *et al.* 2006). Still, it is curious that the random regions had an even lower GC content of 42%. Though the effect size appears to be small, the suggestion that GC content in putative CFSs is higher than expected by chance suggests that the contribution of sequence to fragility need not be rooted in high AT content. The small effect size further suggests that it is likely that the propensity for taking up exogenous DNA is due to a combination of factors.

In addition, this study was able to uncover aphidicolin-sensitive sites that have many of the characteristics of human CFSs, such as high AT content and relatively late replication (Table 3.5). Integration sites such as 100109ins_28, 100109ins_22, and 032509ins_16, which have many of the proposed characteristics of fragility, will warrant close attention in the future.

Factors affecting the insertion assay

In developing the assay, I experimented with many of the methods involved to determine the optimum output. Factors such as amount of input DNA, selection time, cell population heterogeneity, use of splinkerette PCR, and the insert detection scheme were all considered. The amount of DNA used was a key factor; my initial concern was that the amount suggested by the Amaxa kit, 2 µg, might overwhelm the cell and lead to random integration, so I aimed to use as little of the linear insert as possible. The first experiment went as low as 500 ng; in the second iteration of the experiment, I found I was able to identify insert-containing cells that had been exposed to as little as 5 ng of DNA. Detection of the 5 ng sample did not involve selection with hygromycin, but was aided by splinkerette enrichment for insert-containing DNA fragments. It may be that the assay can be implemented with even smaller quantities of DNA. Investigations into the lower threshold of insertion-producing DNA should use some type of selection, such as hygromycin or splinkerette PCR.

The amount of time between transfection and DNA harvest was important to consider, as well. Selection was used to ensure the survival of insert-containing cells, but the longer cells were selected with hygromycin, the more homogenous the population would become. To increase the heterogeneity of the population, I experimented with relatively short growth times of seven days, with no hygromycin treatment. Although there was nothing to give insert-containing cells a selective advantage during the week of growth, and the samples had been transfected with only 5 ng of DNA, I was able to use splinkerette PCR to detect about as many inserts as I had from cells transfected with 100 times as much DNA and selected with hygromycin. The use of a selective agent, such as hygromycin, is therefore not critical to the success of the assay; however, some form of enrichment for insert-containing cells or sequences may be desirable.

Timing, however, is not the only means of affecting insert heterogeneity in the cell population. I experimented with physical separation of the cells to maximize the diversity of the insert pool. If, after transfection, the cells were to be split into as many wells as possible, the overall pool of inserts would be able to avoid becoming completely homogenous, dominated by a few successful cells. By growing cells in twelve small wells instead of one big well, one would expect to increase the number of inserts in the pool by about twelvefold. To date, the only cells that were grown in this way that have been successfully sequenced were those grown in the absence of aphidicolin. This library identified only seven unique inserts, but as the cell did not have aphidicolin to encourage genomic uptake of the construct, it is not surprising that these samples produced fewer inserts than those grown with aphidicolin. Regardless, the reasoning behind the physical separation of cells to encourage heterogeneity is sound; this technique is likely to be useful in maximizing output of the assay.

Splinkerette PCR was instrumental in detecting inserts in a pool that had been transfected with little DNA and not selected with hygromycin. However, the efficiency of enrichment can likely be increased further. The reasons for the initial difficulty of enrichment are not entirely clear, but likely arise from the fact that the insert-containing portion of the extracted S2 DNA is a very small fraction, and discrete banding patterns were not detected after the nested PCR. The approach is effective, as inserts that underwent enrichment via splinkerette PCR were detected with both HTS and TOPO cloning. Still, it is likely that this step could be made more efficient. The main suggested alteration is to use a DNA pool with as many inserts as possible, which can be encouraged by using 500 ng or more starting DNA, selecting with hygromycin, and physically separating the cells. Another way is to take advantage of the relatively high density of insert reads recovered in the splinkerette sample. Despite having less than 3% of the length of the construct to work with, HTS of the splinkerette-enriched 5 ng sample detected more inserts than the 500 ng sample (Figs. 3.2 A & 3.6). The fact that most integrations did not appear to involve an intact construct suggests that insert detection *via* splinkerette PCR could be enhanced by running multiple separate PCR reactions, each using an insert-specific primer situated in a different portion of the insert.

The choice between the use of whole-genome HTS or TOPO cloning depends largely on the time and cost involved. At current efficiencies, TOPO cloning is somewhat less expensive on a per-insert detected basis. In addition, the longer reads available in the sequencing of colony PCR products of TOPO clones give a better chance to identify the location of the insert unambiguously, although the ambiguous integration detected at the histone locus demonstrates that this is not guaranteed. Further, in TOPO cloning, detection of an insert necessarily implies detection of the insert-genome junction. However, the throughput of TOPO clone-based detection is markedly slower, as most TOPO clones of splinkerette PCR products did not lead to insert detection. In addition, while both the HTS and TOPO detection methods would benefit from improvements to integration enrichment *via* splinkerette PCR, HTS has the most to gain. It is conceivable that such improvements could make HTS the more cost-efficient option in the long term.

Materials and Methods

Production of the Insertion Construct

The inserted DNA construct was a modified, linearized version of the pCoHygro vector (Invitrogen, #K4120-01). A QuikChange kit was used to introduce a silent mutation that removed a *Cvi*QI restriction site in the ampicillin resistance gene (Stratagene, #200515). In the iterations of the experiment involving aphidicolin, *gypsy* element insulators were added flanking the hygromycin resistance gene. These insulators were extracted from the pP{RedH-Pelican} vector (GenBank accession #AY342347) by restriction digest and cloned into pCoHygro using the *Hin*dIII site at position 400, and the *Eco*ICRI site at position 2287. The construct was linearized by cutting with the restriction enzyme *Bsp*QI (NEB, #R0712S) for three hours, and gel purifying the product using Buffer QG (Qiagen, #19063) and a PureLink PCR purification kit (Invitrogen, #K3100-01). The ends of the construct were made blunt by klenow (NEB, #M0210S).

Integration of DNA

pCoHygro was transfected into *Drosophila* S2 cells using the Amaxa Nucleofector and the Cell Line Nucleofector Kit V (Lonza, VCA-1003). Cells were provided by members of the Rodgers laboratory. Cells were maintained in Sf-900 II serum-free media (Invitrogen, #10902096) with 1x antibiotic-antimycotic (Invitrogen, #15240062), and kept in T25 and T75 flasks (BD Falcon, #353014 and #353135, respectively). Quantification of cells was done by counting 10 µl of a 1:10 dilution in a hemocytometer.

4.5×10^6 S2 cells were used for each transfection. Aphidicolin was dissolved in dimethyl sulfoxide (DMSO) to a 2 mg/ml concentration; a 40 µM working stock was made by diluting 1.35 µl of the main stock in 200 µl of DMSO (Aphidicolin: Acros Organics, 1 mg, #611970010; DMSO, Sigma-Aldrich, 100 ml, #154938). If aphidicolin was to be used in the treatment, it was added to the cells at a concentration of 0.4 µM, and allowed to sit for 30 minutes. Cells were passaged to 1.5 µl tubes, and spun for 5 minutes at 2,000 rpm (381xG). The media was removed, and each tube was resuspended in 90 µl Solution V from the Nucleofector kit. About 10 µl of either 5 or 500 ng of pCoHygro DNA, depending on the experiment, was added to the cells. Cells were moved to a Nucleofector cuvette and electroporated using the S2 cells setting of the Amaxa machine. Cells were then passaged to media in 6- or 24-well plates, depending on the experiment; if the sample had been treated with aphidicolin, this media also contained 0.4 µM aphidicolin. In some cases, a single cuvette worth of electroporated cells was split between 12 wells of a 24-well plate.

Transfected cells were grown until they had achieved greater than 75% confluence in the well; they were then passaged either to a T25 flask, or to another tissue culture plate. Selection with hygormycin (Roche, #10-843-555-001) began when the cells again reached confluence, starting at 5 µl/ml and gradually increasing to 10 µl/ml. Selection continued until cells reached >95% confluence in 10 µl/ml hygromycin; this took 15-19 days, depending on the experiment. If cells were not treated with hygromycin, cells were grown for one week. At the end of either time period, cells were harvested and their DNA extracted.

Extraction of S2 cell DNA

Cells were harvested by pipetting, and spun down in a 1.5 ml tube at for 5 minutes at 500xG. The supernatant was removed, the pellet resuspended in 1 ml ice-cold PBS, and the tube spun 5 minutes at 500xG. After discarding the supernatant, cells were resuspended in 425 µl of digest buffer (100 mM NaCl; 10 mM Tris-HCl, pH 8.0; 25 mM EDTA, pH 8.0; 0.5% (w/v) SDS; 0.1 mg/ml freshly-added proteinase K). Samples were incubated at 50° C for 12-18 hours. Phenol/chloroform/isoamyl alcohol extraction was performed at least twice to purify the DNA.

Insert detection by high-throughput sequencing

Illumina library construction was performed using Illumina paired-end adapters (part #1001782) as described in the paired-end sample preparation guide (#1005063 Rev. D), with the following differences. First, 5 µg DNA was fragmented via sonication using a Bioruptor (Diagenode, #UCD-200), and concentrated with the PureLink PCR purification kit (Invitrogen) prior to end repair. End repair took place in a 50 ml reaction volume. After the post-end repair column purification, size selection took place: the DNA was run on a 2% agarose gel for 60 minutes, and DNA in the 350-400 bp range was extracted and purified. During adapter ligation, 2 µl of adapter was used in a 30 µl reaction volume. A second gel extraction on a 2% agarose gel was run as described in the original protocol. The PCR to enrich adapter-modified DNA was carried out with 0.5 µl iProof high-fidelity polymerase (Bio-Rad, #172-5301), rather than Phusion polymerase. The primers used were PAGE-purified oligos ordered from Integrated DNA Technologies, with the same sequence as those provided in the Illumina kit.

In some cases, insert-containing DNA fragments were enriched via splinkerette PCR (UREN *et al.* 2009). This was initiated by ligating blunt-end splinkerette adapters to the DNA fragments after the end-repair step. A splinkerette primer (CGAAGAGTAACCGTTGCTAGGAGAGACC) and insert-specific primer (GGGGCGGAGCCTATGGAAAA) were used to amplify insert-containing fragments (35 cycles of PCR, 56° C annealing temperature, 1 minute extension); the products were further amplified by nested PCR (splinkerette: AGACTGGTGTCGACACTAGTGG; insert-specific: TTTGCTGGCCTTTTGCTC; 35 cycles of PCR, 58° C annealing temperature, 1 minute extension). Primers were designed using Primer3 (ROZEN and SKALETSKY 2000). At this point, the PCR product was either used to resume Illumina library prep by proceeding to the addition of Illumina PE adapters, or used for TOPO cloning.

Libraries were sequenced on the Illumina Genome Analyzer IIx for 36-76 cycles. Reads were mapped to the fly genome (release 5) and to the insert sequence using Bowtie (LANGMEAD *et al.* 2009); inserts were detected by filtering for clusters that had one sequencing read containing at least 30 nt of insert DNA, and a paired read with genomic DNA sequence.

Insert detection by TOPO cloning

Library preparation was begun as above, but after the addition of adenosine to the ends of each fragment, the DNA was subjected to cloning as per the TOPO TA cloning kit (Invitrogen, #K4500-01). A 30 second heat shock was used to transform TOP10 *E. coli* cells, and colonies were grown for ~16 hours on kanamycin-containing agarose plates. Colonies were analyzed by using M13 primers for colony PCR (forward primer, GTAAAACGACGGCCAG; reverse primer, CAGGAAACAGCTATGAC); the PCR products were also sequenced using the M13 primers.

Sequencing of insert-genome junctions

When sequencing the junctions of inserts detected by HTS, I used the sequence information obtained from the GA IIx to design primers for PCR amplification. Sequences were amplified for 35 cycles, with a 60.9° C annealing temperature and a 30 second extension, and sequenced via Sanger sequencing.

Integration environment analysis

The regions surrounding integration sites were analyzed with the Galaxy framework (BLANKENBERG *et al.* 2010; GOECKS *et al.* 2010). Much of the analysis relied on data tracks from the modENCODE project (CELNIKER *et al.* 2009). Insert intervals were compared to equally-sized intervals of random location; the random data was produced using the random number generator at random.org (HAAHR 1998). Copy number variation was taken into account when generating random regions (ZHANG *et al.* 2010). Statistical analyses were performed using the InStat software (GraphPad Software). An unpaired t-test was used in cases in which the sample did not differ significantly from a normal distribution; the Mann-Whitney test was performed if this was not the case.

Chapter IV

General Discussion and Future Directions

I have reported that CFSs are present in *D. melanogaster*, and have identified their location. In doing so, I have explored the connection between induced and natural CFSs, and provided evidence of the connection between DmBLM and *Drosophila* CFSs. My genome-wide assay has identified putative fragile regions at a high resolution. In addition, I have produced two versatile assays that may be applied to future studies of CFSs, or to DNA breakage in general. Here, I provide further interpretation of the ramifications of my results, and propose future experiments to build on my findings.

Evolutionary conservation

The CFSs I have detected in *D. melanogaster* is in keeping with suggestions of evolutionary conservation implicated by studies that have detected fragile regions in budding yeast (ADMIRE *et al.* 2006; LEMOINE *et al.* 2005). The reasons for the conservation of such a potentially deleterious genomic feature are unclear. One possible explanation is that some useful genomic features may be necessarily fragile. For example, many micro RNAs are located at human CFSs; the selective advantage provided by the micro RNAs may outweigh any detrimental effects of fragility (CALIN *et al.* 2004).

Another possibility is that CFSs could be used for large-scale genomic rearrangements, such as translocations and inversions, and that a certain amount of these are necessary for the adaptability and survival of a species. However, the mouse orthologs of specific human CFSs do not correlate with sites of evolutionary DNA rearrangements between the two species (HELMRICH *et al.* 2006; HELMRICH *et al.* 2007). Still, it may be that some CFSs involved in a large-scale rearrangement lose the characteristics that made them fragile, thus preventing one from finding an association between synteny breaks and CFSs. To determine if this is the case, one would need to examine not only CFSs orthologous to known human CFSs, but also CFSs that do not necessarily have a human ortholog.

If it turns out that CFSs are used for evolutionary large-scale rearrangements, this would suggest that either less derived organisms should have more CFSs than the more derived, or that new CFSs appear throughout evolution. The fact that most of the CFSs detected have been described in humans could be seen as evidence for the second possibility, but it is important to keep in mind that human CFSs have received the most study. If similar effort were put toward detecting and analyzing the CFSs of less derived organisms, such that we could be confident that we had a catalog of most or all CFSs in a given species, it would be easier to make an equivalent comparison. The work presented here is a step in such a direction.

A recent paper found that sites of recurrent evolutionary breaks, rather than functional constraints on chromosome breakage, are the most parsimonious explanation for the evolution of *Drosophila* chromosomes (VON GROTTHUSS *et al.* 2010). The authors make reference to “fragile regions”, which they intend to refer to sites prone to breakpoint reuse in genome rearrangements on an evolutionary timescale, not the short-term fragility seen in RFSs and CFSs. Still, as more is learned about *Drosophila* CFSs, it will be interesting to see if a meaningful relationship exists between CFSs and evolutionary breakpoints.

Differences in the crossover and integration assays

My crossover assay and my insertion assay provide different pictures of the fragility of *2L*. For example, the crossover assay indicates the distal tip of *2L* to be the most fragile part of the chromosome arm, but the insertion assay does not have any integrations in that region. However, this serves to underscore the relevant differences between the two experiments. The first is that tissue matters. The crossover assay detects DSBs that takes place in the pre-meiotic male germline. The insertion assay, on the other hand, uses S2 cells, which have the characteristics of macrophages (SCHNEIDER 1972). Tissue-specificity of fragility has been reported in human cells, and has been suggested to be related to variations in replication timing profiles between cell types (LETESSIER *et al.* 2011). Therefore, it may simply be that regions that are fragile in the male germline are not fragile in macrophages, and *vice versa*. The aneuploidy of the S2 cells could also potentially affect these results (ZHANG *et al.* 2010).

The second relevant difference between the assays is the status of DmBLM. Mutations in *mus309* were necessary in the crossover assay, as they drove the formation of the mitotic crossovers I mapped; there is no indication of mutated DmBLM in S2 cells. Human BLM has been associated with CFSs, though, and DmBLM may play an active role in *Drosophila* CFS stability. Human cells that lack BLM exhibit delayed replication, and BLM has been implicated in maintaining the integrity of stalled replication forks (DAVIES *et al.* 2007; RASSOOL *et al.* 2003). Furthermore, BLM associates with ultrafine anaphase bridges (CHAN *et al.* 2007), which may represent the entangled DNA of replicating chromosomes. These bridges have been detected at CFSs (CHAN *et al.* 2009). The overall implication is that BLM helps the cell recover from replication stress; in the absence of BLM, cells experience DNA breaks in the regions of chromosomes that are either unreplicated or contain DNA entangled with another chromosome. In short, BLM may be involved in preventing breaks at CFSs, so its absence could affect the distribution of CFSs observed.

This issue lends itself well to future experiments. Knockdown of genes in S2 cells via RNA interference is relatively straightforward (CLEMENS *et al.* 2000). It should be possible, then, to knock down DmBLM in the cells while performing the integration assay. If integrations on *2L* in the absence of DmBLM matched the distribution produced by the crossover assay, it would suggest that DmBLM played a larger role than tissue specificity in influencing fragility. If there were no change in the *2L* integrations, it would suggest that tissue specificity played a larger role; intermediate results would indicate a combination of the two factors as influencing fragility. This experiment could also be done with or without aphidicolin to investigate possible synergistic effects.

Investigation of the physical location of DmBLM may serve as a useful future experiment, as well. As noted above, the ultrafine bridges associated with human BLM are also associated with many CFSs (CHAN *et al.* 2007; CHAN *et al.* 2009). Of importance is the fact that the bridges have been interpreted to be the entangled DNA of sister chromatids which require BLM to decatenate. This interpretation, combined with the association with CFSs, means that the bridges may represent the regions of DNA that the cell has the greatest difficulty replicating. Precise localization of the bridges could thus be used to find the most fragile regions within a CFS. A high-resolution way to examine such sites would be chromatin immunoprecipitation followed by HTS (ChIP-seq) (ROBERTSON *et al.* 2007). One would crosslink proteins to DNA, shear the DNA, pull down DNA associated with DmBLM using an antibody, reverse the crosslinks, and use HTS to identify the sequences of the collected fragments. These sequences would indicate where DmBLM had been bound at the time of the crosslinking, thus identifying the locations of the ultrafine bridges at a high resolution. As such bridges have been detected in unperturbed human cells, the experiment could be done with or without treatment by agents like aphidicolin, providing another opportunity to examine the relationship between natural and induced CFSs. Of course, such an experiment could also be carried out in human cell culture with an antibody for human BLM. Versions of the experiment that pull down other proteins associated with the ultrafine bridges, such as FANCD2, are also possible (CHAN *et al.* 2009).

Use of Eureqa for future studies

The causes of fragility, and their relative contribution to the fragility of a given site, remain to be determined. I believe a useful way to approach this is through application of the Eureqa software (SCHMIDT and LIPSON 2009). This program was designed to identify mathematical relationships between sets of data, and to do so in a way that non-trivial correlations are emphasized. It does this by calculating equations that not only fit the data, but the partial derivatives of the equation also fit the derivatives of pairs of the data. The software has been used to derive the calculations of natural laws; for example, data on the motion of a pendulum is sufficient for Eureqa to derive Force = mass * acceleration (SCHMIDT and LIPSON 2009).

I propose to apply this technique to CFSs, to determine the relative contribution of various factors to making a region fragile. One would have to determine a numerical way of representing fragility; initially, this would be easiest to accomplish with human data. The fragility of specific CFSs relative to other CFSs is fairly well-established for the most fragile CFSs in humans (DENISON *et al.* 2003). One could then provide Eureqa with the data on the fragility of various sites, along with factors such as replication timing and numerical representations of primary sequence characteristics, and anything else the user felt pertinent. The utility of the equation produced by Eureqa would be to provide an estimate of the relative contribution of the different factors toward fragility. In the pendulum example above, one can see that mass and acceleration contribute equally to determining the resulting force. Such an approach should be able to guide future investigations into the nature of fragility, and provides a complementary approach to other multifactor analyses, such as multiple regression.

Direct tests of fragility

Future experiments should also directly test the characteristics implicated in fragility. Two good ways to do this would be to introduce fragility into a normally stable region, and to remove fragility from a fragile region. Increased resolution from SNP mapping in the CO assay, or a greater number of integrations in the S2 cell assay, should help identify the relevant characteristics. To introduce fragility, one would create a *de novo* CFS – not derived from any specific *Drosophila* sequence, but having the characteristics of a minimal fragile region. This site could then be inserted into specific genomic locations using ΦC31-mediated integration and tested for fragility with my crossover assay (BATEMAN *et al.* 2006). Alternatively, adding a selectable marker and 1-2 kb of specific *Drosophila* sequence to either end of the construct and transfecting it into S2 cells would establish a line of cells with the new CFS stably integrated in a targeted manner, which could in turn be used to test for fragility in the integration assay (RONG and GOLIC 2000). Experiments in which a section of human FRA3B was inserted in ectopic regions of the genome – and found to maintain fragility – suggest that this type of experiment would be feasible with a *de novo* CFS (RAGLAND *et al.* 2008).

A complementary approach to these would be the ablation of fragility in known CFSs. Here, the approach will depend on the characteristic of fragility one wishes to abolish. For example, replication timing and distance to origins could be disrupted by mutation of nearby origins. Alternatively, deletions flanking the CFS could be used to bring origins closer. Direct manipulation of the sequence, such as removal of a minimal fragile region, could be used to test the contribution of sequence to fragility. Indeed, it has been found that such deletions in CFSs can reduce fragility (ARLT *et al.* 2002). However, such deletions do not always result in a complete reduction in fragility, and sometimes do not result in a reduction of fragility at all (CORBIN *et al.* 2002; DURKIN *et al.* 2008) . Furthermore, it has been pointed out that simply deleting sequence conflates the contribution of primary sequence and origin distance; taking out putative fragile sequence could also result in a reduction of the distance between origins, thereby making it easier for the cell to complete replication of the region before mitosis (LETESSIER *et al.* 2011). Therefore, I propose that future experiments that seek to test the influence of fragile sequence replace putative fragile sequence with a non-fragile sequence of equal length. Such a replacement could be accomplished through gene targeting by homologous recombination (RONG and GOLIC 2000).

Useful modifications to assays

Both assays described here are extremely versatile, and have the potential for applications beyond the detection and analysis of *Drosophila* CFSs. For example, the cell-based integration assay is relatively easy to port to human cell culture. Modifications may be required for the type of selection used, but the main scheme of transfecting cells with a selectable construct, with or without aphidicolin, and using paired-end HTS to identify integration sites, should work well in human cells. Others have used integration-based approaches to clone CFSs of known location, but have done so on a relatively small scale, analyzing only a single fragile site at a time (MISHMAR *et al.* 1998; RASSOOL *et al.* 1996). The whole-genome HTS integrated into my insertion assay allows for a much broader scope, while retaining a high resolution.

The ability to study human CFSs with a high throughput will become increasingly important. Seventy-five of the eighty-eight human CFSs have not been studied at the molecular level; my integration assay could be used to provide an in-depth characterization of such sites (LUKUSA and FRYNS 2008). In addition, even among the thirteen CFSs that have been molecularly characterized, studies have indicated that the differences in CFS expression between individuals, or between different tissues, are substantial enough to warrant further study (DENISON *et al.* 2003; LETESSIER *et al.* 2011). Much of the study of CFSs has taken place in lymphocytes, so many tissues remain to be studied. Using my integration assay in human cells provides an efficient means of examining different tissues and patient samples.

There are opportunities to expand the scope of damage detected by my assays, as well. The integration assay was used here to identify insertions into sites of both endogenous damage and damage induced by aphidicolin. However, the exogenous DNA damaging agents examined by the assay need not be limited to aphidicolin. For example, one could treat cells with ionizing radiation, and locate the resulting DSBs at a very high resolution. To my knowledge, there is little understood about which genomic regions, if any, that might be susceptible to ionizing radiation; the assay I have developed provides the opportunity to investigate the issue. A complementary approach might be to treat flies with IR, and use the crossover assay to ascertain the location of breaks.

Conclusion

My discovery of CFSs in *Drosophila* has identified an important link in the evolutionary conservation of CFSs, and added to a growing body of evidence that connects BLM and fragile sites. In addition, I developed two useful assays, and laid the groundwork for future studies. The SNP mapping of mitotic crossovers – reciprocal and otherwise – and the ability to take the HTS integration assay in new directions will be important steps to determining the causes of genome instability. As noted above, these future experiments need not be limited to *Drosophila*; studies of various human populations and tissues can benefit from this work, as well as examinations of DNA damaging agents. In total, this work has contributed to a growing body of knowledge on chromosome fragility, and puts us in an excellent position to learn even more.

Appendix Chapter

Transcription Initiation from Within *P* Elements Generates Hypomorphic Mutations in *Drosophila melanogaster*

Preface

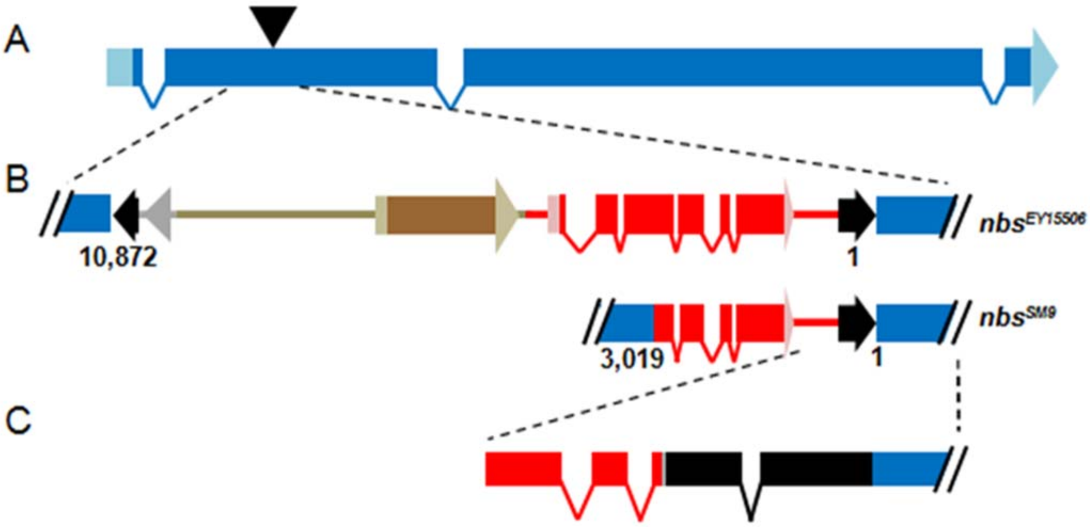
The following is a manuscript written by myself and Dr. Jeff Sekelsky, which was submitted to the journal *Genetics* as a Note. While not directly related to CFSs, this work speaks to the effect that DNA insertions can have on their local genomic environment. At the time of the submission of this dissertation, the manuscript is under review at *Genetics*.

Results & Discussion

Genetically engineered *P* elements have been used as tools for experimental manipulation of the Drosophila genome (BELLEN *et al.* 2004; THIBAULT *et al.* 2004). These elements tend to insert near transcription start sites, frequently resulting in hypomorphic mutations (SPRADLING *et al.* 1995). This result is sometimes surprising, as insertions within early exons might be expected to produce null mutations rather than hypomorphs. Here, we report that transcription initiation from within a *P* element may be responsible for some mutations being hypomorphic rather than null.

We first observed this phenomenon in a *P* insertion allele of *nbs* (Figure 5.1A). The $nbs^{EY15506}$ allele contains a 10.9 kb *P{EPgy2}* element inserted into the coding sequence in the second exon. Although the insertion separates two conserved domains, it creates not a null allele, but a hypomorphic, separation-of-function allele. The nbs^{SM9} derivative has an internal deletion of about 7 kb of this *P* element (Figure 5.1B), but retains the hypomorphic nature of $nbs^{EY15506}$.

We previously reported that the hypomorphic character of *nbs*^{SM9} was due, at least in part, to transcription that initiated from within the *P* element (Figure 5.1C) (MUKHERJEE *et al.* 2009). Using 5' RACE, we found that transcription initiated in the region downstream of the 3' end of the *white* (*w*) gene carried on the *P* element – a region that is located 3' to the endogenous *w* locus, but is not part of the annotated *w* gene span. In addition, we detected that three introns, with canonical splice donor and acceptor sequences, had been spliced out (Figure 5.1C, D). Two of these introns are located in the region downstream of *w*, but the third is within the *P* element 5' end, which is transcribed in the opposite direction to *P* transposase. The terminal three bp of the *P* end form an in-frame start codon. Results from our assays suggested that retention of some NBS functions was due to production of a protein lacking the N-terminal FHA domain.



D

```
AACTTACACACCTGCATACCTTACATCAAAAACTCGTTTATCGCTACATAAAACACCGGGATAT
ATTTTTTATATACATACTTTTCAAATCGCGCCCTCTTCATAATTCACCTCCACCACACCACG
TTTCGTAGTTGCTCTTTCGCTGTCTCCCACCCGCTCTCCGCAACACATTCACCTTTTGTTCGAC
GACCTTGGAGCGACTGTCGTTAGTTCCGCGCGATTCGGTTCGCTCAAATGGTTCCGAGTGGTTC
ATTTCGTCTCAATAGAAATTAGTAATAAATATTTgtatgtacaatttatttgctccaatatatt
tgtatatatttccctcacagctatatttattctaatttaatattatgactttttaagGTAATTT
TTTGTGACCTGTTCGGAGTGATTAGCGTTACAATTTGAACTGAAAGTGACATCCAGTGTTTGTT
CCTTGTGTAGATGCATCTCAAAAAAATGgtgggcataatagtgttgtttatatatatcaaaaat
aacaactataataataagaatacatttaatttagAAAATGCTTGGATTTCACTGGAACTAGAAT
TAATTCGGCTGCTGCTCTAAACGACGCATTTCGTACTCCAAAGTACGAATTTTTTCCCTCAAGC
TCTTATTTTCATTAAACAATGAACAGGACCTAACGCACAGTCACGTTATTGTTTACATAAATGA
TTTTTTTTACTATTCAAACTTACTCTGTTTGTGTACTCCCACTGGTATAGCCTTCTTTTATCTT
TTCTGGTTCAGGCTCTATCACTTTACTAGgtacggcatctgcgttgagtcgcctccttttaaat
gtctgaccttttgcagGTGCAGCCTTCCACTGCGAATCATTAAAGTGGGTATCACAAATTTGGG
AGTTTTCACCAAGGCTGCACCCAAGGCTCTGCTCCCACAATTTTCTCTTAATAGCACACTTCGG
CACGTGAATTAATTTTACTCCAGTCACAGCTTTGCAGCAAAATTTGCAATATTTCATTTTTTTT
TATTCCACGTAAGGGTTAATGTTTTCAAAAAAAAATTCGTCCGCACACAACCTTTCCTCTCAAC
AAGCAAACGTGCACTGAATTTAAGTGTATACTTCGGTAAGCTTCGGCTATCGACGGGACCACCT
TATGTTATTTCATCATG...
```

Figure 5.1. A cryptic transcript initiates from within a *P* element in *nbs*. (A) The *nbs*^{EY15506} allele has a 10.9 kb *P{EPgy2}* element inserted into the 2nd exon. The *P* element is oriented with the 5' end nearest the 3' end of the *nbs* coding region. Lighter-colored regions indicate UTRs. (B) The *P* element contains the *w* gene and its surrounding endogenous regions (red) oriented toward the 5' *P* end (black), as well as the *yellow* gene (brown) and a *S. cerevisiae UAS* promoter (gray arrow). The *nbs*^{SM9} allele is an internally deleted version of *nbs*^{EY15506} that retains about 3 kb of the *P* element. (C) The cryptic transcript detected by the 5´ RACE System for Rapid Amplification of cDNA Ends, Version 2.0 (Invitrogen; catalog #18374-058). Transcription initiation appears to initiate in the region downstream of *w*. Three introns are spliced out in adult RNA samples. (D) The sequence of the cryptic transcript is represented in capital letters; lowercase letters represent sequences present in the DNA that are spliced out in the transcript. Red letters, the region downstream of *w*; gray letters, *P* element backbone sequence; black letters, the 5' *P* end.

The nbs^{SM9} allele is unusual in that the *P* element is inserted into coding sequences. To determine whether insertions in other regions can also produce hypomorphic alleles due to transcription from a *P* element, we analyzed an insertion into the 5' untranslated region (UTR) of the *Slbp* gene (Figure 5.2A). This insertion creates a hypomorphic allele (SULLIVAN *et al.* 2001). The structure of the *P{EP}* element in this allele differs from that of the *P{EPgy2}*element we analyzed in *nbs*, but in both cases the element is inserted so that the *w* gene is transcribed in the same direction as the gene into which the *P* element is inserted (Figure 5.2B). SULLIVAN *et al*. (2001) carried out a *P* excision screen and generated hypomorphic and null alleles, both caused by internal deletions within the *P* element, without deletion of *Slbp* sequence. Deletions that removed the region downstream of the *w* gene, such as the null allele $Slbp^{15}$, do not produce *Slbp* transcript. In contrast, the hypomorphic allele $Slbp^{10}$, which lost much of the *P* element but retained the region downstream of *w*, produces both *Slbp* transcript and SLBP protein.

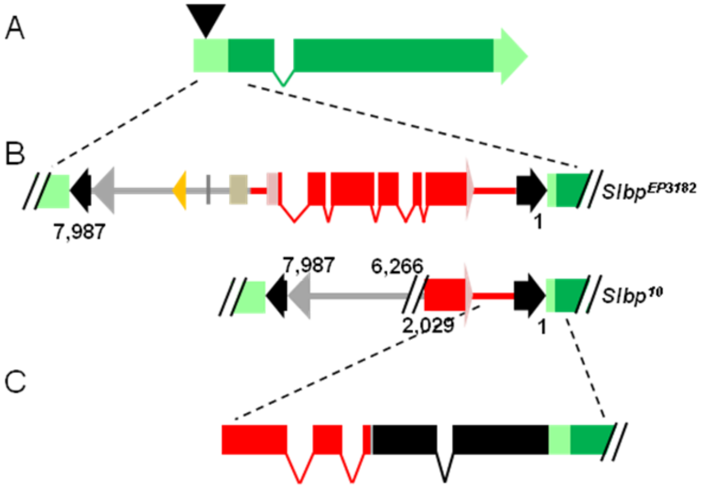


Figure 5.2. A similar cryptic transcript initiates from within a *P* element in *Slbp*. (A) The $Slbp^{EP3182}$ allele has an 8 kb *P{EP}* element inserted in the 5' UTR. The *P* element is oriented with the 5' end nearest the *Slbp* coding region. (B) The *P{EP}* element contains the *w* gene and its surrounding endogenous regions (red) oriented toward the 5' *P* end (black); other features are as described by Rørth (1996). $Slbp^{10}$ has a deletion from 2,029-6,266. (C) The cryptic transcript was detected by using a combination of PCR and the 5′ RACE System. Our PCR did not facilitate the detection of the 5' end of the transcript, but the position of the upstream primer is consistent with transcription initiation that occurs in the region downstream of *w*. Three cryptic introns are spliced out in adult RNA samples; a larval sample did not have the intron in the 5' *P* end spliced out.

We performed PCR on cDNA from *Slbp*^{10}/*TM6B* adults and 2nd instar larvae, and found that there is also transcription from the region downstream of *w* in this *P{EP}* element. In adults, the sequence and splice sites are identical to those we detected in *nbs*^{SM9} (Figure 5.1D; Figure 5.2C). Intriguingly, removal of the intron in the *P* end was not detected in cDNA isolated from larvae, raising the possibility that splicing of the intra-*P* transcript may be regulated by tissue and/or developmental timing.

These findings provide insight into the nature of hypomorphic alleles that arise from certain transposable element insertions. For example, there is no guarantee that transcription from within the transposable element will match the expression pattern of the gene's native promoter. If this is the case, altered expression may result in different severities, potentially ranging from wild-type function in some tissues to complete absence of function in other tissues.

Features of these cryptic transcripts that affect translation efficiency may play a role in the reduced function as well. While *Slbp*^{10} mutants produce SLBP protein, they do so at reduced levels (SULLIVAN *et al.* 2001). This may be due to characteristics of the cryptic transcript. Among the exons on the *P*-element-*Slbp* transcript, the exon in which translation begins – a fusion of the 5' *P* end and the remaining sequence of the 5' UTR and start of the coding region of *Slbp* – is the only exon that does not contain a short open reading frame. This suggests that translation can successfully take place only when ribosomes bind to that key exon, as binding to the upstream exons would lead to premature termination of translation, and potentially to nonsense-mediated decay (NMD) (GATFIELD *et al.* 2003; PELTZ *et al.* 1993). This feature may contribute to the hypomorphic character of some mutants with intra-*P* transcription: even if transcript were being produced at the normal rate, transcript degradation may ensure that not all transcripts would survive to be translated, leading to a reduced level of protein. The cryptic transcript of *nbs*^{SM9} does not share this characteristic. Rather, we found that levels of *nbs*^{SM9} transcript appeared to be similar to that of wild-type (MUKHERJEE *et al.* 2009). This suggests that NMD may not be involved in reducing the levels of *nbs* transcript. This is curious, as the *P* elements in *Slbp*^{10} and *nbs*^{SM9} have identical sequence in the region of the cryptic transcript. This suggests that the stability of these transcripts is affected by additional factors that are not immediately apparent.

The position and orientation of a *P* element insertion may have additional ramifications for mutagenesis. Insertions in introns may invoke alternative splicing of the native transcript, possibly resulting in impaired functionality of the transcript or protein. Additionally, it may be that insertions that take place in a gene span, but which transcribe the region downstream of *w* in the opposite direction as the gene, may produce antisense RNA. These transcripts could initiate RNA interference of the native transcript, providing another means of producing a hypomorph.

The region from which intra-*P* transcription appears to initiate corresponds to the genomic location *X*:2,683,995..2,684,631, directly upstream of the gene *CG32795* (*Drosophila melanogaster* genome release 5.34) (TWEEDIE *et al.* 2009). However, stable transcripts from this region have not been detected. Transcripts in this region are absent from most modENCODE genome browser RNA expression profiling tracks, except in those of immunoprecipitation of Argonaute-1 or 2 in S2 cells (CELNIKER *et al.* 2009). It may be that this region does undergo transcription, but as no translation follows, the transcripts are not stable or are degraded. This phenomenon may be analogous to that of cryptic unstable transcripts observed in budding yeast (WYERS *et al.* 2005).

The *w* gene is a common marker in engineered transposable elements in Drosophila. Transcription initiation from within these elements is likely to contribute to the hypomorphic mutations generated by many of these insertion alleles.

References

ADAMS, M. D., M. MCVEY and J. J. SEKELSKY, 2003 Drosophila BLM in double-strand break repair by synthesis-dependent strand annealing. Science **299:** 265-267.

ADMIRE, A., L. SHANKS, N. DANZL, M. WANG, U. WEIER *et al.*, 2006 Cycles of chromosome instability are associated with a fragile site and are increased by defects in DNA replication and checkpoint controls in yeast. Genes Dev **20:** 159-173.

ARLT, M. F., and T. W. GLOVER, 2010 Inhibition of topoisomerase I prevents chromosome breakage at common fragile sites. DNA Repair (Amst) **9:** 678-689.

ARLT, M. F., D. E. MILLER, D. G. BEER and T. W. GLOVER, 2002 Molecular characterization of FRAXB and comparative common fragile site instability in cancer cells. Genes Chromosomes Cancer **33:** 82-92.

BACHRATI, C. Z., and I. D. HICKSON, 2003 RecQ helicases: suppressors of tumorigenesis and premature aging. Biochem J **374:** 577-606.

BATEMAN, J. R., A. M. LEE and C. T. WU, 2006 Site-specific transformation of Drosophila via phiC31 integrase-mediated cassette exchange. Genetics **173:** 769-777.

BECKER, N. A., E. C. THORLAND, S. R. DENISON, L. A. PHILLIPS and D. I. SMITH, 2002 Evidence that instability within the FRA3B region extends four megabases. Oncogene **21:** 8713-8722.

BELLEN, H. J., R. W. LEVIS, G. LIAO, Y. HE, J. W. CARLSON *et al.*, 2004 The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes. Genetics **167:** 761-781.

BLANKENBERG, D., G. VON KUSTER, N. CORAOR, G. ANANDA, R. LAZARUS *et al.*, 2010 Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol **Chapter 19:** Unit 19 10 11-21.

BURROW, A. A., L. E. WILLIAMS, L. C. PIERCE and Y. H. WANG, 2009 Over half of breakpoints in gene pairs involved in cancer-specific recurrent translocations are mapped to human chromosomal fragile sites. BMC Genomics **10:** 59.

CALIN, G. A., C. SEVIGNANI, C. D. DUMITRU, T. HYSLOP, E. NOCH *et al.*, 2004 Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A **101:** 2999-3004.

CARTER, B. S., C. M. EWING, W. S. WARD, B. F. TREIGER, T. W. AALDERS *et al.*, 1990 Allelic loss of chromosomes 16q and 10q in human prostate cancer. Proc Natl Acad Sci U S A **87:** 8751-8755.

CASPER, A. M., P. NGHIEM, M. F. ARLT and T. W. GLOVER, 2002 ATR regulates fragile site stability. Cell **111:** 779-789.

CELNIKER, S. E., L. A. DILLON, M. B. GERSTEIN, K. C. GUNSALUS, S. HENIKOFF *et al.*, 2009 Unlocking the secrets of the genome. Nature **459:** 927-930.

CHA, R. S., and N. KLECKNER, 2002 ATR homolog Mec1 promotes fork progression, thus averting breaks in replication slow zones. Science **297:** 602-606.

CHAGANTI, R. S., S. SCHONBERG and J. GERMAN, 1974 A manyfold increase in sister chromatid exchanges in Bloom's syndrome lymphocytes. Proc Natl Acad Sci U S A **71:** 4508-4512.

CHAN, K. L., P. S. NORTH and I. D. HICKSON, 2007 BLM is required for faithful chromosome segregation and its localization defines a class of ultrafine anaphase bridges. Embo J **26:** 3397-3409.

CHAN, K. L., T. PALMAI-PALLAG, S. YING and I. D. HICKSON, 2009 Replication stress induces sister-chromatid bridging at fragile site loci in mitosis. Nat Cell Biol **11:** 753-760.

CHEN, T., A. SAHIN and C. M. ALDAZ, 1996 Deletion map of chromosome 16q in ductal carcinoma in situ of the breast: refining a putative tumor suppressor gene region. Cancer Res **56:** 5605-5609.

CHENG, C. H., and R. D. KUCHTA, 1993 DNA polymerase epsilon: aphidicolin inhibition and the relationship between polymerase and exonuclease activity. Biochemistry **32:** 8568-8574.

CIRULLI, E. T., R. M. KLIMAN and M. A. NOOR, 2007 Fine-scale crossover rate heterogeneity in Drosophila pseudoobscura. J Mol Evol **64:** 129-135.

CLEMENS, J. C., C. A. WORBY, N. SIMONSON-LEFF, M. MUDA, T. MAEHAMA *et al.*, 2000 Use of double-stranded RNA interference in Drosophila cell lines to dissect signal transduction pathways. Proc Natl Acad Sci U S A **97:** 6499-6503.

COQUELLE, A., E. PIPIRAS, F. TOLEDO, G. BUTTIN and M. DEBATISSE, 1997 Expression of fragile sites triggers intrachromosomal mammalian gene amplification and sets boundaries to early amplicons. Cell **89:** 215-225.

CORBIN, S., M. E. NEILLY, R. ESPINOSA, 3RD, E. M. DAVIS, T. W. MCKEITHAN *et al.*, 2002 Identification of unstable sequences within the common fragile site at 3p14.2: implications for the mechanism of deletions within fragile histidine triad gene/common fragile site at 3p14.2 in tumors. Cancer Res **62:** 3477-3484.

DAVIES, S. L., P. S. NORTH and I. D. HICKSON, 2007 Role for BLM in replication-fork restart and suppression of origin firing after replicative stress. Nat Struct Mol Biol **14:** 677-679.

DE BONT, R., and N. VAN LAREBEKE, 2004 Endogenous DNA damage in humans: a review of quantitative data. Mutagenesis **19:** 169-185.

DENISON, S. R., R. K. SIMPER and I. F. GREENBAUM, 2003 How common are common fragile sites in humans: interindividual variation in the distribution of aphidicolin-induced fragile sites. Cytogenet Genome Res **101:** 8-16.

DEVON, R. S., D. J. PORTEOUS and A. J. BROOKES, 1995 Splinkerettes--improved vectorettes for greater efficiency in PCR walking. Nucleic Acids Res **23:** 1644-1645.

DHILLON, V. S., S. A. HUSAIN and G. N. RAY, 2003 Expression of aphidicolin-induced fragile sites and their relationship between genetic susceptibility in breast cancer, ovarian cancer, and non-small-cell lung cancer patients. Teratog Carcinog Mutagen **Suppl 1:** 35-45.

DILLON, L. W., A. A. BURROW and Y. H. WANG, 2010 DNA instability at chromosomal fragile sites in cancer. Curr Genomics **11:** 326-337.

DURKIN, S. G., M. F. ARLT, N. G. HOWLETT and T. W. GLOVER, 2006 Depletion of CHK1, but not CHK2, induces chromosomal instability and breaks at common fragile sites. Oncogene **25:** 4381-4388.

DURKIN, S. G., and T. W. GLOVER, 2007 Chromosome fragile sites. Annu Rev Genet **41:** 169-192.

DURKIN, S. G., R. L. RAGLAND, M. F. ARLT, J. G. MULLE, S. T. WARREN *et al.*, 2008 Replication stress induces tumor-like microdeletions in FHIT/FRA3B. Proc Natl Acad Sci U S A **105:** 246-251.

EATON, M. L., J. A. PRINZ, H. K. MACALPINE, G. TRETYAKOV, P. V. KHARCHENKO *et al.*, 2011 Chromatin signatures of the Drosophila replication program. Genome Res **21:** 164-174.

FREUDENREICH, C. H., 2005 Molecular Mechanisms of Chromosome Fragility. ChemTracks-Biochemistry and Molecular Biology **18**.

FUNDIA, A., N. GORLA and I. LARRIPA, 1995 Non-random distribution of spontaneous chromosome aberrations in two Bloom Syndrome patients. Hereditas **122:** 239-243.

GARDINER, K., 1995 Human genome organization. Curr Opin Genet Dev **5:** 315-322.

GATFIELD, D., L. UNTERHOLZNER, F. D. CICCARELLI, P. BORK and E. IZAURRALDE, 2003 Nonsense-mediated mRNA decay in Drosophila: at the intersection of the yeast and mammalian pathways. Embo J **22:** 3960-3970.

GLOOR, G. B., C. R. PRESTON, D. M. JOHNSON-SCHLITZ, N. A. NASSIF, R. W. PHILLIS *et al.*, 1993 Type I repressors of P element mobility. Genetics **135:** 81-95.

GLOVER, T. W., 1981 FUdR induction of the X chromosome fragile site: evidence for the mechanism of folic acid and thymidine inhibition. Am J Hum Genet **33:** 234-242.

GLOVER, T. W., C. BERGER, J. COYLE and B. ECHO, 1984 DNA polymerase alpha inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes. Hum Genet **67:** 136-142.

GLOVER, T. W., and C. K. STEIN, 1987 Induction of sister chromatid exchanges at common fragile sites. Am J Hum Genet **41:** 882-890.

GOECKS, J., A. NEKRUTENKO and J. TAYLOR, 2010 Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol **11:** R86.

GOSCIN, L. P., and J. J. BYRNES, 1982 DNA polymerase delta: one polypeptide, two activities. Biochemistry **21:** 2513-2518.

HAAHR, M., 1998 http://www.random.org/.

HANDT, O., G. R. SUTHERLAND and R. I. RICHARDS, 2000 Fragile sites and minisatellite repeat instability. Mol Genet Metab **70:** 99-105.

HAVIV-CHESNER, A., Y. KOBAYASHI, A. GABRIEL and M. KUPIEC, 2007 Capture of linear fragments at a double-strand break in yeast. Nucleic Acids Res **35:** 5192-5202.

HELMRICH, A., K. STOUT-WEIDER, K. HERMANN, E. SCHROCK and T. HEIDEN, 2006 Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes. Genome Res **16:** 1222-1230.

HELMRICH, A., K. STOUT-WEIDER, A. MATTHAEI, K. HERMANN, T. HEIDEN *et al.*, 2007 Identification of the human/mouse syntenic common fragile site FRA7K/Fra12C1--relation of FRA7K and other human common fragile sites on chromosome 7 to evolutionary breakpoints. Int J Cancer **120:** 48-54.

HIBI, K., T. TAKAHASHI, K. YAMAKAWA, R. UEDA, Y. SEKIDO *et al.*, 1992 Three distinct regions involved in 3p deletion in human lung cancer. Oncogene **7:** 445-449.

HIGUCHI, K., T. KATAYAMA, S. IWAI, M. HIDAKA, T. HORIUCHI *et al.*, 2003 Fate of DNA replication fork encountering a single DNA lesion during oriC plasmid DNA replication in vitro. Genes Cells **8:** 437-449.

HIRSCH, B., 1991 Sister chromatid exchanges are preferentially induced at expressed and nonexpressed common fragile sites. Hum Genet **87:** 302-306.

HUEBNER, K., and C. M. CROCE, 2003 Cancer and the FRA3B/FHIT fragile locus: it's a HIT. Br J Cancer **88:** 1501-1506.

IKEGAMI, S., T. TAGUCHI, M. OHASHI, M. OGURO, H. NAGANO *et al.*, 1978 Aphidicolin prevents mitotic cell division by interfering with the activity of DNA polymerase-alpha. Nature **275:** 458-460.

JIANG, Y., I. LUCAS, D. J. YOUNG, E. M. DAVIS, T. KARRISON *et al.*, 2009 Common fragile sites are characterized by histone hypoacetylation. Hum Mol Genet **18:** 4501-4512.

KAROW, J. K., A. CONSTANTINOU, J. L. LI, S. C. WEST and I. D. HICKSON, 2000 The Bloom's syndrome gene product promotes branch migration of holliday junctions. Proc Natl Acad Sci U S A **97:** 6504-6508.

LAFAVE, M. C., and J. SEKELSKY, 2009 Mitotic recombination: why? when? how? where? PLoS Genet **5:** e1000411.

LANGMEAD, B., C. TRAPNELL, M. POP and S. L. SALZBERG, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol **10:** R25.

LAROCQUE, J. R., D. L. DOUGHERTY, S. K. HUSSAIN and J. SEKELSKY, 2007a Reducing DNA polymerase alpha in the absence of Drosophila ATR leads to P53-dependent apoptosis and developmental defects. Genetics **176:** 1441-1451.

LAROCQUE, J. R., B. JAKLEVIC, T. T. SU and J. SEKELSKY, 2007b Drosophila ATR in double-strand break repair. Genetics **175:** 1023-1033.

LE BEAU, M. M., F. V. RASSOOL, M. E. NEILLY, R. ESPINOSA, 3RD, T. W. GLOVER *et al.*, 1998 Replication of a common fragile site, FRA3B, occurs late in S phase and is delayed further upon induction: implications for the mechanism of fragile site induction. Hum Mol Genet **7:** 755-761.

LEE, P. S., P. W. GREENWELL, M. DOMINSKA, M. GAWEL, M. HAMILTON *et al.*, 2009 A fine-structure map of spontaneous mitotic crossovers in the yeast Saccharomyces cerevisiae. PLoS Genet **5:** e1000410.

LEMOINE, F. J., N. P. DEGTYAREVA, K. LOBACHEV and T. D. PETES, 2005 Chromosomal translocations in yeast induced by low levels of DNA polymerase a model for chromosome fragile sites. Cell **120:** 587-598.

LETESSIER, A., G. A. MILLOT, S. KOUNDRIOUKOFF, A. M. LACHAGES, N. VOGT *et al.*, 2011 Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. Nature **470:** 120-123.

LUKUSA, T., and J. P. FRYNS, 2008 Human chromosome fragility. Biochim Biophys Acta **1779:** 3-16.

MANKOURI, H. W., and I. D. HICKSON, 2007 The RecQ helicase-topoisomerase III-Rmi1 complex: a DNA structure-specific 'dissolvasome'? Trends Biochem Sci **32:** 538-546.

MATZNER, I., L. SAVELYEVA and M. SCHWAB, 2003 Preferential integration of a transfected marker gene into spontaneously expressed fragile sites of a breast cancer cell line. Cancer Lett **189:** 207-219.

MCBRIDE, G., 1979 Fragile X chromosome related to mental retardation in males. Jama **242:** 1829-1830.

MCVEY, M., S. L. ANDERSEN, Y. BROZE and J. SEKELSKY, 2007 Multiple functions of Drosophila BLM helicase in maintenance of genome stability. Genetics **176:** 1979-1992.

MCVEY, M., J. R. LAROCQUE, M. D. ADAMS and J. J. SEKELSKY, 2004 Formation of deletions during double-strand break repair in Drosophila DmBlm mutants occurs after strand invasion. Proc Natl Acad Sci U S A **101:** 15694-15699.

MCVEY, M., and S. E. LEE, 2008 MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet **24:** 529-538.

MISHMAR, D., A. RAHAT, S. W. SCHERER, G. NYAKATURA, B. HINZMANN *et al.*, 1998 Molecular characterization of a common fragile site (FRA7H) on human chromosome 7 by the cloning of a simian virus 40 integration site. Proc Natl Acad Sci U S A **95:** 8141-8146.

MOORE, J. K., and J. E. HABER, 1996 Capture of retrotransposon DNA at the sites of chromosomal double-strand breaks. Nature **383:** 644-646.

MUKHERJEE, S., M. C. LAFAVE and J. SEKELSKY, 2009 DNA damage responses in Drosophila nbs mutants with reduced or altered NBS function. DNA Repair (Amst) **8:** 803-812.

NEGRINI, M., C. MONACO, I. VORECHOVSKY, M. OHTA, T. DRUCK *et al.*, 1996 The FHIT gene at 3p14.2 is abnormal in breast carcinomas. Cancer Res **56:** 3173-3179.

O'KEEFE, L. V., and R. I. RICHARDS, 2006 Common chromosomal fragile sites and cancer: focus on FRA16D. Cancer Lett **232:** 37-47.

PALAKODETI, A., Y. HAN, Y. JIANG and M. M. LE BEAU, 2004 The role of late/slow replication of the FRA16D in common fragile site induction. Genes Chromosomes Cancer **39:** 71-76.

PELTZ, S. W., A. H. BROWN and A. JACOBSON, 1993 mRNA destabilization triggered by premature translational termination depends on at least three cis-acting sequence elements and one trans-acting factor. Genes Dev **7:** 1737-1754.

POPESCU, N. C., D. ZIMONJIC and J. A. DIPAOLO, 1990 Viral integration, fragile sites, and proto-oncogenes in human neoplasia. Hum Genet **84:** 383-386.

RAGLAND, R. L., M. W. GLYNN, M. F. ARLT and T. W. GLOVER, 2008 Stably transfected common fragile site sequences exhibit instability at ectopic sites. Genes Chromosomes Cancer **47:** 860-872.

RASSOOL, F. V., M. M. LE BEAU, M. L. SHEN, M. E. NEILLY, R. ESPINOSA, 3RD *et al.*, 1996 Direct cloning of DNA sequences from the common fragile site region at chromosome band 3p14.2. Genomics **35:** 109-117.

RASSOOL, F. V., T. W. MCKEITHAN, M. E. NEILLY, E. VAN MELLE, R. ESPINOSA, 3RD *et al.*, 1991 Preferential integration of marker DNA into the chromosomal fragile site at 3p14: an approach to cloning fragile sites. Proc Natl Acad Sci U S A **88:** 6657-6661.

RASSOOL, F. V., P. S. NORTH, G. J. MUFTI and I. D. HICKSON, 2003 Constitutive DNA damage is linked to DNA replication abnormalities in Bloom's syndrome cells. Oncogene **22:** 8749-8757.

ROBERTSON, G., M. HIRST, M. BAINBRIDGE, M. BILENKY, Y. ZHAO *et al.*, 2007 Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods **4:** 651-657.

RONG, Y. S., and K. G. GOLIC, 2000 Gene targeting by homologous recombination in Drosophila. Science **288:** 2013-2018.

RORTH, P., 1996 A modular misexpression screen in Drosophila detecting tissue-specific phenotypes. Proc Natl Acad Sci U S A **93:** 12418-12422.

ROZEN, S., and H. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol **132:** 365-386.

SANCAR, A., L. A. LINDSEY-BOLTZ, K. UNSAL-KACMAZ and S. LINN, 2004 Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. Annu Rev Biochem **73:** 39-85.

SARAI, A., J. MAZUR, R. NUSSINOV and R. L. JERNIGAN, 1989 Sequence dependence of DNA conformational flexibility. Biochemistry **28:** 7842-7849.

SCHMIDT, M., and H. LIPSON, 2009 Distilling free-form natural laws from experimental data. Science **324:** 81-85.

SCHNEIDER, I., 1972 Cell lines derived from late embryonic stages of Drosophila melanogaster. J Embryol Exp Morphol **27:** 353-365.

SCHWARTZ, M., E. ZLOTORYNSKI, M. GOLDBERG, E. OZERI, A. RAHAT *et al.*, 2005 Homologous recombination and nonhomologous end-joining repair pathways regulate fragile site stability. Genes Dev **19:** 2715-2726.

SCHWARTZ, M., E. ZLOTORYNSKI and B. KEREM, 2006 The molecular basis of common and rare fragile sites. Cancer Lett **232:** 13-26.

SHRIDHAR, R., V. SHRIDHAR, X. WANG, W. PARADEE, M. DUGAN *et al.*, 1996 Frequent breakpoints in the 3p14.2 fragile site, FRA3B, in pancreatic tumors. Cancer Res **56:** 4347-4350.

SPRADLING, A. C., D. M. STERN, I. KISS, J. ROOTE, T. LAVERTY *et al.*, 1995 Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. Proc Natl Acad Sci U S A **92:** 10824-10830.

STONE, D. M., P. B. JACKY, D. D. HANCOCK and D. J. PRIEUR, 1991 Chromosomal fragile site expression in dogs: I. Breed specific differences. Am J Med Genet **40:** 214-222.

STONE, D. M., K. E. STEPHENS and J. DOLES, 1993 Folate-sensitive and aphidicolin-inducible fragile sites are expressed in the genome of the domestic cat. Cancer Genet Cytogenet **65:** 130-134.

SULLIVAN, E., C. SANTIAGO, E. D. PARKER, Z. DOMINSKI, X. YANG *et al.*, 2001 Drosophila stem loop binding protein coordinates accumulation of mature histone mRNA with cell cycle progression. Genes Dev **15:** 173-187.

SUTHERLAND, G. R., E. BAKER and R. I. RICHARDS, 1998 Fragile sites still breaking. Trends Genet **14:** 501-506.

SUTHERLAND, G. R., M. I. PARSLOW and E. BAKER, 1985 New classes of common fragile sites induced by 5-azacytidine and bromodeoxyuridine. Hum Genet **69:** 233-237.

TEDESCHI, B., P. VERNOLE, M. L. SANNA and B. NICOLETTI, 1992 Population cytogenetics of aphidicolin-induced fragile sites. Hum Genet **89:** 543-547.

THIBAULT, S. T., M. A. SINGER, W. Y. MIYAZAKI, B. MILASH, N. A. DOMPE *et al.*, 2004 A complementary transposon tool kit for Drosophila melanogaster using P and piggyBac. Nat Genet **36:** 283-287.

THORLAND, E. C., S. L. MYERS, D. H. PERSING, G. SARKAR, R. M. MCGOVERN *et al.*, 2000 Human papillomavirus type 16 integrations in cervical tumors frequently occur in common fragile sites. Cancer Res **60:** 5916-5921.

TWEEDIE, S., M. ASHBURNER, K. FALLS, P. LEYLAND, P. MCQUILTON *et al.*, 2009 FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res **37:** D555-559.

UREN, A. G., H. MIKKERS, J. KOOL, L. VAN DER WEYDEN, A. H. LUND *et al.*, 2009 A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. Nat Protoc **4:** 789-798.

VERKERK, A. J., M. PIERETTI, J. S. SUTCLIFFE, Y. H. FU, D. P. KUHL *et al.*, 1991 Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell **65:** 905-914.

VILENCHIK, M. M., and A. G. KNUDSON, 2003 Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer. Proc Natl Acad Sci U S A **100:** 12871-12876.

VON GROTTHUSS, M., M. ASHBURNER and J. M. RANZ, 2010 Fragile regions and not functional constraints predominate in shaping gene organization in the genus Drosophila. Genome Res **20:** 1084-1096.

WAN, C., A. KULKARNI and Y. H. WANG, 2010 ATR preferentially interacts with common fragile site FRA3B and the binding requires its kinase activity in response to aphidicolin treatment. Mutat Res **686:** 39-46.

WILKE, C. M., B. K. HALL, A. HOGE, W. PARADEE, D. I. SMITH *et al.*, 1996 FRA3B extends over a broad region and contains a spontaneous HPV16 integration site: direct evidence for the coincidence of viral integration sites and fragile sites. Hum Mol Genet **5:** 187-195.

WOODRUFF, R. C., and J. N. THOMPSON, 1977 An analysis of spontaneous recombination in Drosophila melanogaster males. Heredity **38:** 291-307.

WYERS, F., M. ROUGEMAILLE, G. BADIS, J. C. ROUSSELLE, M. E. DUFOUR *et al.*, 2005 Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell **121:** 725-737.

YUNIS, J. J., and A. L. SORENG, 1984 Constitutive fragile sites and cancer. Science **226:** 1199-1204.

ZHANG, H., and C. H. FREUDENREICH, 2007 An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in S. cerevisiae. Mol Cell **27:** 367-379.

ZHANG, Y., J. H. MALONE, S. K. POWELL, V. PERIWAL, E. SPANA *et al.*, 2010 Expression in aneuploid Drosophila S2 cells. PLoS Biol **8:** e1000320.