Application of Whole Genome Sequencing and Bioinformatics Tools to Investigate Novel Interactions in DNA Repair in *Drosophila melanogaster*

A dissertation submitted by

Barbara Taylor Sands-Marcinkowski

In partial fulfillment of the requirements for the degree of

Master of Science

In

Biology

TUFTS UNIVERSITY

February 2018

Advisor: Mitch McVey

ABSTRACT

An organism's genome is constantly exposed to endogenous and exogenous forms of damage. If left unrepaired, DNA damage can pose a threat to the regulation and function of genes located near the site of the lesion. To address the multiple types of DNA damage that accumulate on a daily basis, organisms have evolved specific pathways to deal with each one. Although most of these pathways have been studied for multiple years or even decades, there are still many interactions and mechanisms that have yet to be elucidated. In recent years, the study of genomes and their maintenance has changed significantly thanks to drastic improvements to DNA sequencing technologies and the corresponding data analysis tools. In the work described here, whole genome sequencing (WGS) and bioinformatic data analysis were used to address two research questions, each of which focused on investigating DNA repair proteins and their roles in DNA repair and genomic stability in *Drosophila melanogaster*. First, I investigated the cause of a synthetic larval lethality, which was discovered after creating double mutants for two conserved helicases: DmBlm and DmHelQ. Second, I analyzed mutants obtained from an ethyl-methane sulfonate (EMS) mutagenesis screen that was conducted in a trans-lesion synthesis (TLS)-deficient background, in order to elucidate genes and interactions involved in the template-switching (TS) pathway of DNA damage tolerance (DDT). For both projects, I applied a single bioinformatic pipeline to WGS data, which led to the identification of strong candidates for causative variants in each.

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my advisor, Mitch McVey, for his unwavering and unparalleled support, advice, and encouragement throughout my three and a half years as a student and researcher here at Tufts. Mitch accepted me into the McVey Lab family, and made it possible for me to take my prior experience with DNA library preparation, and my interest in sequencing and biotechnology, and apply it to two different research projects. Mitch taught me a great deal about research, molecular biology, and DNA repair; but he also provided me with a prime example of what it means to be a great teacher, communicator, and leader. When I came to a crossroads, and decided on a new path for my future as a scientist, Mitch supported me throughout the entire transition and never failed to recognize and encourage my potential, for which I am both humbled and eternally grateful. I do not know how I would have made it to this point with his support and guidance, and I can probably never thank him enough for all he has done for me as a scientist and as a young professional.

I would also like to thank Sergei Mirkin, who met with me before I even applied to Tufts, and gave me a spot to rotate in his laboratory once I was accepted. As a strong leader and an expert in his field, Sergei has always been a huge inspiration to me, and I am so grateful that he has continued to provide advice (on both science and life) and support all throughout my time here at Tufts. His class on DNA Structure and Function was an awesome and fun learning experience for me, as well. His suggestions and input on my projects have been invaluable, and my committee could not be complete without him.

Thanks also go out to the other two members of my committee: Steve Fuchs and Erik Dopman. Steve's class (BIO-243) was a great way to start my graduate career at Tufts. It provided a crash-course in reading and presenting that benefited me immensely in the years to follow. I asked Erik to be on my committee so that he might contribute his perspective and experience with sequencing and bioinformatic analysis approaches. At every committee meeting, Erik had brilliant suggestions for my projects, and his ideas were always pushing me to think harder about the basic assumptions that we make as scientists, and how to acknowledge and address them. Both Steve and Erik have been invaluable pillars on my committee, and I thank them for all of their advice and suggestions throughout the past few years. Other Tufts folks to whom thanks are due: everyone who keeps the Biology Department running and makes it the awesome department that it is, and Charlie Sykes, not only for being an awesome mentor to my husband, but also for helping me receive the Tufts Provost Fellowship for my first two years at Tufts.

I have so many amazing friends that I want to acknowledge; because through ups and downs, happy times and stressful times, and everything in between; they were there for me. Special shout-outs within the Tufts Community go to everyone I met through the Graduate Student Council, and all of the biology graduate students (especially my main girl, Cassandra Donatelli). Another special shout out goes to the McVey lab undergrads, graduate students, and post-docs, past and present. Our lab is the best lab, the most fun lab, and has been like a family for my entire time here. To Terrence, Tokio, Justin, and Mai: keep it real

for me. To Alice: keep them in line. I'm going to miss all of you so much! Oh, and don't forget Lab Spirit Week! Outside of Tufts, I want to thank everyone from #TeamMarcinkowski (all 18 of you and your significant others), plus Andrew and Alyssa, Nikki, Ellis, Katie, and my Cornell Andrew: you are all my dearest friends, and I cannot imagine my life without any of you in it. (P.S. Stasia: I love you, bestie).

I want to thank my family members, both immediate and extended, who have been incredibly supportive of me since Day 1. To my parents, Biff and Brenda: thank you for encouraging me and believing in me throughout my entire academic and career journey like no one else could. To my brother, Benjamin, and his girlfriend, Emily, thank you guys for being the most awesome friends and for always being there to hang out and chat with me about grad school, or to help me escape from the stress of grad school, depending on what was needed. Thank you to my in-laws: Susan and Andy; Ryan, Kate, Mia, and Robert; and Dana and Alex. I couldn't have married into a more awesome, loving family if I had tried.

Last but not least, thanks to my most treasured family member and my best friend: my husband, Matthew. We've made it through 8 months of long-distance and now we're on to a new adventure, and I can't wait! Your constant support, encouragement, wisdom, and shared laughter made my graduate school journey possible. I can't imagine having done this without you on my team, cheering me on. I love you so much, and every day I feel blessed to be your wife.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BSA = Bulk Segregant Analysis
BWA = Burrows-Wheeler Alignment
DDT = DNA Damage Tolerance
dHJ = double Holliday Junction
D-loop = Displacement loop
DNA = Deoxyribonucleic acid
dNTP = DeoxyriboNucleotide triphosphate
DSB = Double Strand Break
DSBR = Double Strand Break Repair
EMS = Ethyl-Methane Sulfonate
FGS = Forward Genetic Screen
GATK = Genome Analysis Tool Kit
HDR = Homology Directed Repair
HJ = Holliday Junction
HR = Homologous Recombination
HRR = Homologous Recombination Repair
HU = HydroxyUrea
ICL = Interstrand CrossLink
MMS = Methyl-Methane Sulfonate
NGS = Next Generation Sequencing
NTP = Nucleotide triphosphate
SDSA = Synthesis Dependent Strand Annealing
SNV = Single Nucleotide Variation
TLS = Trans-Lesion Synthesis
TS = Template Switching
VCF = Variant Call Format
VEP = Variant Effect Predictor
VRM = VCF Recombination Mapping
WES = Whole Exome Sequencing
WGS = Whole Genome Sequencing

Application of Whole Genome Sequencing and Bioinformatics Tools to Investigate Novel Interactions in DNA Repair in *Drosophila melanogaster*

Chapter 1: Introduction

Variant Discovery and Analysis in the WGS World: Applying a Sequencing and Bioinformatics Framework to Two Classical Genetic Studies in *Drosophila melanogaster*

ABSTRACT

Both the benefits and the potential of DNA sequencing technologies are apparent across multiple fields of biological research, as well as in industry and the medical field. Sequencing has also allowed the more traditional techniques and assays employed in these fields to be approached in innovative ways, and among the most prominent benefactors are studies in forward genetics. The traditional methods used to connect genetic variants with observed phenotypes proved both prolonged and laborious. Now, comparative next generation genomics provides significant relief. Whole genome sequencing (WGS) has already been applied to genetic screens in multiple systems, and while approaches for both sequencing and data analyses have been evaluated and compared, comprehensive start-to-finish frameworks for research labs transitioning to the use of WGS have not been extensively described. Here, I provide a brief history of WGS, give examples of its impacts in classical forward genetics, and present my own example of a start-to-finish WGS framework built for *Drosophila melanogaster*. I both obtained and analyzed WGS data, and identified candidate mutations, without traditional genetic mapping methods. I applied the workflow to two different forward genetics studies, which addressed unique biological questions related to DNA repair and genomic stability. In parallel, I took steps to implement best practices for on-site data storage, organization, and manipulation. I propose that this framework could be maintained, adjusted, and reapplied for similar forward genetics experiments in *Drosophila melanogaster*, and with some reconfiguration and specification, to those in other model systems.

A brief history of the sequencing revolution in molecular biology

Since its introduction, DNA sequencing has provided ongoing advantages to endeavors in both applied and investigative biology. One of the first and most impactful results of the technology was the ability to, over time, piece together nucleic acid sequences that constituted entire organismal genomes. Frederick Sanger and colleagues developed and applied perhaps the most well-known early method of sequencing, which was achieved by radiolabeling deoxyribonucleotide triphosphate (dNTP) monomers and incorporating them on DNA restriction fragment scaffolds. Labeled fragments were then separated and visualized on polyacrylamide gels. Once primers were associated with the restriction fragments, the maximum length of the read was only 150-200 base pairs (bp), at best, and this limited the length and overall amount of DNA that could be sequenced at a time (Sanger, Air, et al., 1977). The Sanger sequencing method was first applied to the approximately 5.4 kilobase (kb) long genome of the bacteriophage ϕX174, also known as PhiX, and this genome is still used as a sequencing standard in many labs to date (Sanger, Air, et al., 1977). The introduction of chain-terminating dNTPs provided additional speed and accuracy to the technique (Sanger, Nicklen, & Coulson, 1977), and further improvements in labeling and detection methods were made in the years to follow. One of the modifications of Sanger sequencing led to the development of Shotgun sequencing, which was based on sequencing a collection of larger, random genome fragments in the form of a library of bacterial clones (Sanger, Coulson, Hong, Hill, & Petersen, 1982; Weber & Myers, 1997). The assembly of a genome from these fragments was

Comment [FSM1]: This is a very late reference if you're talking about the history of sequencing. I believe shotgun sequencing started in the early 80s

Comment [BS2]: Added original ref

based on overlaps in sequence information at the fragment ends. Subsequently, cosmids and yeast artificial chromosomes were used to accommodate even larger DNA fragments (Cooper, 2000). Together, these advances prompted the publishing of many model organism genomes just a few decades after the initial development of Sanger sequencing. These genomes included bacteria (*Haemophilus influenza* and *Escherichia coli* K-12), archaea (*Methanococcus jannaschii*), baker's yeast (*Saccharomyces cerevisiae*), nematode worms (*Caenorhabditis elegans*), Arabidopsis (*Arabidopsis thaliana*), fruit flies (*Drosophila melanogaster*), humans (*Homo sapiens*), and mice (*Mus musculus*) (M D Adams et al., 2000; Arabidopsis Genome Initiative, 2000; Blattner, 1997; Bult et al., 1996; Fleischmann et al., 1995; Lander et al., 2001; The C. elegans Sequencing Consortium, 1998; Venter et al., 2001; Waterston et al., 2002; Williams et al., 1996). Access to full genomes afforded researchers working in these model systems with a boon of new potential in genetic manipulation, conservation studies, gene discovery, and more. However, for smaller private and academic research labs, sequencing the genomes of individual specimens on a continual basis was still a relatively inaccessible endeavor due to high costs, read length restrictions, and lack of automation.

The next major improvement to sequencing technology, which over time systematically addressed the major limitations of Sanger sequencing, was next-generation sequencing (NGS): a set of sequencing platforms characterized by massively parallelized, ultra-high-throughput sample processing. A number of biotechnology companies emerged with different versions of next generation

sequencing technology: 454 Life Sciences/Roche, Solexa/Illumina, Applied Biosystems, Helicos Biosciences, and Pacific Biosciences were among the biggest competitors in the early years of NGS, and are reviewed in detail elsewhere (Mardis, 2008; Metzker, 2010; Shendure & Ji, 2008; Shendure, Mitra, Varma, & Church, 2004). While these companies all used slightly different versions of sequencing biochemistry to accomplish the rapid, massively parallelized reaction, most followed an underlying formula. The more abundant version of the reaction chemistry, generally, began as follows: DNA was first fragmented and associated with index/barcode sequences which allow the fragment to anneal to a primer bound to a bead (454/Roche, Applied Biosystems) or glass slide (Solexa/Illumina, Helicos) and in some cases also included primer-binding sites for downstream amplification of the fragment library. Next, bound fragments were amplified so that thousands of copies of the template sequence would be available for the sequencing reaction. In the case of the Helicos platform, an amplification step was not included, as the sequencing reaction took place on single molecules. Finally, the sequencing reaction itself took place via cycles of binding and then washing off reversible chain-terminating dNTPs with fluorescent labels. The signals from these fluorescent labels were read by a camera and then translated to an output signal with each cycle, in order to determine the identity of the base at that position.

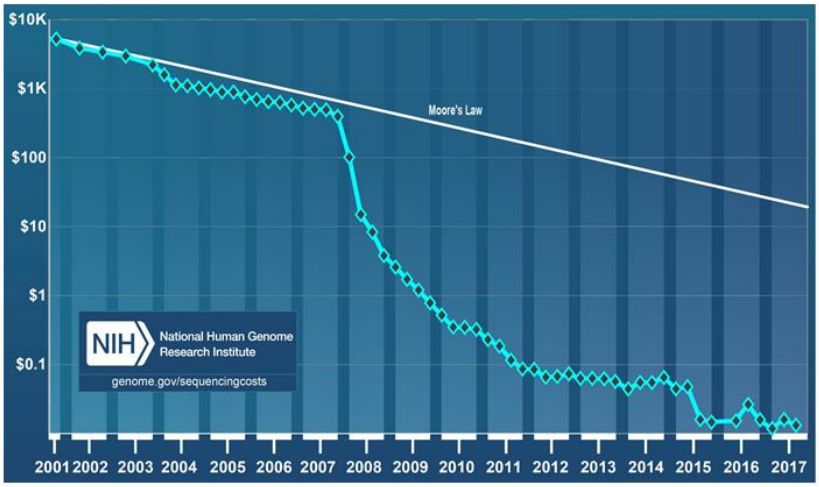
Other platforms, such as Ion Torrent, which was purchased by Life Technologies in 2010 (MacArthur, 2010), and single molecular real-time (SMRT) sequencing by Pacific Biosciences, boasted more unique reaction chemistries. Ion

Torrent used the release of protons during the incorporation of nucleotides to detect sequencing information as synthesis took place. SMRT immobilized polymerases in micro-wells to systematically detect fluorescent labels released from dNTPs as they were incorporated from a single molecule template (Quail et al., 2012). More recently, Oxford Nanopore announced ultra-long reads, low cost, and rapid results with their nanopore sequencing platform. By passing an ion current through individual nanopore channels within a chip-based array, they could detect changes in the current as a DNA molecule was pulled through (Huang, Romero-Ruiz, Castell, Bayley, & Wallace, 2015). Each base combination passing through the nanopore would create a different signature in the current output. Although less competitive in the early days of NGS technology, Ion Torrent, SMRT, and Nanopore have been developed extensively in recent years, and are now actively competing with the prevailing industry giant, Illumina.

Despite various approaches to reaction chemistry, all of the NGS technologies share in the accomplishment of allowing scientists to sequence the entire genome of an organism all at once. Furthermore, NGS not only made WGS readily attainable, but also made it affordable. According to data collected by the NHGRI Genome Sequencing Program since 2001, the arrival of NGS began to drive the cost-per-base of sequencing the human genome down drastically between 2007 and 2008 (**Figure 1-1a**). This change corresponded with a noticeable and favorable deviation from trend estimates derived from Moore's Law, which represented estimates of cost-per-base assuming only a biannual

doubling in computational capability (Wetterstrand, 2016). By the end of 2015, the cost of sequencing an entire human genome approached the important

a.



b.

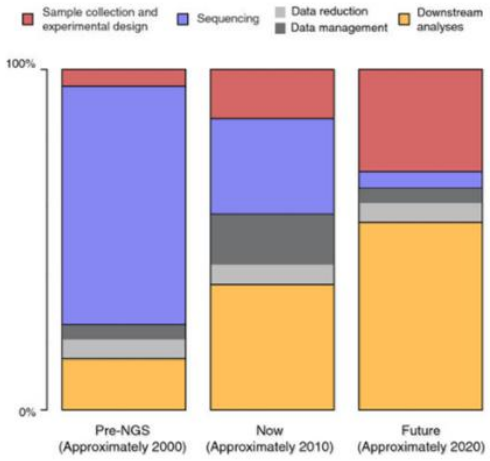


Figure 1-1. Lower sequencing costs are accompanied by shifts in spending. (a) Data from the NHGRI are shown. The black points and aqua trend line represent the cost of sequencing per raw megabase of human genome (log scale) and how it has changed over the years. This is compared to the white trend line, which represents estimates of sequencing costs obtained via Moore's Law. (b) The percentage of sequencing costs attributable to various processes is shown. While the cost of sequencing has dropped dramatically in accordance with (a), the cost of experimental design and downstream analysis, in particular, have increased. 1-(b) modified from Sboner, Mu, Greenbaum, Auerbach, & Gerstein, 2011.

benchmark price point of $1000 (National Human Genome Research Institute, 2016), and currently some prices have dipped below this point, although there is much debate surrounding published and actual costs, and many estimates are based on whole exome sequencing (WES) rather than WGS. Sample preparation, labor, and required data analysis are additional complicating factors in cost analyses (**Figure 1-1b**) (Sboner et al., 2011; Van Nimwegen et al., 2016). Nonetheless, the increased affordability of sequencing reactions that accompanied NGS technologies provided newfound accessibility for a larger population of scientists in both the public and private sectors. This expansion allows us to apply WGS to both new and pre-existing research questions.

WGS in forward genetics: a modern upgrade for classical techniques

The term "forward genetics" is often used interchangeably with "classical genetics". This is because the forward genetics approach—seeking the underlying genetic cause of an observed phenotype—is one that was central in genetics research long before the arrival of sequencing and other modern molecular techniques. One of the most common manifestations of the forward genetics approach is the forward genetic screen (FGS). In screens, the genetic material of an organism is exposed to a mutagen, which causes changes—usually on the single nucleotide level—in random locations along the chromosome. The

frequency of these changes depends on the dosage of mutagen applied at the outset. Usually, a mutagen-treated organism is crossed to a wild-type organism, and changes that have occurred in the germline of the mutagen-treated organism can be passed to offspring. The generation at which phenotypes are screened for depends on the nature of the underlying mutation, and mutant strains obtained from screens are usually maintained in a heterozygous state. Non-lethal mutations functioning dominantly will have phenotypes that are visible in the first generation (F_1). Alternatively, to screen for recessively functioning mutations, additional generations are necessary. Individual F_1 organisms are mated to wild type organisms, followed by interbreeding of F_2 organisms, and screening for phenotypes in the third generation (F_3), where the mutation can be obtained in the homozygous state (Kile & Hilton, 2005; St Johnston, 2002). Once an interesting phenotype is identified, the hunt to track down the underlying genomic change begins. Forward genetics studies can also seek causes for de novo phenotypes occurring outside of the context of a screen. In other words, these would be spontaneous changes to the genome that were not induced with a mutagen, and have instead appeared randomly over time. De novo changes present an additional challenge; the change that causes the underlying cause cannot be assigned to a consequence typical of a certain mutagen. For example, alkylating mutagens like ethyl-methane sulfonate (EMS) tend to cause transition mutations, from GC to AT and from AT to GC (Sega, 1984). In a forward genetics approach, the ability to use parameters like these can go a long way in trimming the list of candidates for causative changes. However, in the case of spontaneous phenotypes, all types of

single nucleotide changes as well as insertion and deletion events will also warrant consideration.

Before scientists had access to WGS, the process of identifying a region of interest for causal mutations took months or even years, depending on the model system. Usually, multiple traditional mapping methods were employed in this effort, and often included crosses of mutants stocks to each other and/or to stocks with large chromosomal deficiencies in order to detect a lack of complementation, indicating a mutation within the same region or within the tested deficiency, respectively. Mutant stocks could also be crossed to mapping stocks with visible markers tiled along the genome, and interbreeding of the F1 generation would yield offspring that could be genotyped for these markers in order to narrow down a region of interest via traditional linkage mapping analysis. For mutations in heterochromatic regions, this process had to be extended, and required even finer scale mapping (Moresco, Li, & Beutler, 2013). Finally, any final candidates would have to be tested and confirmed as causal for the observed phenotype. Overall, the traditional mapping methods required to identify causal mutation candidates from screens created a bottleneck in the utility of forward genetics endeavors in the lab. The power of WGS in forward genetics lies in its ability to remove this bottleneck, and in some cases, to eliminate traditional mapping methods entirely from the process.

Only a few years after most NGS platforms were first introduced, biologists were predicting its impacts on forward genetics. As early as 2008,

scientists predicted that WGS would replace Sanger sequencing of amplified DNA when it came to both fine-scale mapping and mutation discovery in genetic screens (Mardis, 2008). Just one year later, a research group used the Illumina platform to apply WGS to mutants obtained via EMS mutagenesis in *D. melanogaster* (Blumenstiel et al., 2009). In the same year, a method called SHOREmap was published, which computationally accomplished genomic mapping using WGS data obtained from large pools of recombinant individuals. This method also simultaneously identified variant calls using the same WGS data set (Schneeberger et al., 2009). Additionally, the SHOREmap team identified that their system could be applied to bulk segregant analyses (BSA), in which large pools of individuals displaying a phenotype are compared to a pool of wild-type individuals in order to find loci that segregate at 100% in the phenotype pool and 0% in the wild-type pool. However, these and many other early examples of the application of WGS in forward genetics still relied heavily on some amount of traditional genetic mapping and/or the use of hundreds of recombinants.

In 2010, another method was published which took isolates obtained in a mutagenesis screen in *C. elegans* and backcrossed them to unmutagenized starter stocks (Zuryn, Le Gras, Jamet, & Jarriault, 2010). Progeny were then screened for the phenotype of interest. These progeny, when sequenced, contained a cluster of EMS-induced mutations that harbored the putative causal mutation. While this approach may have expedited the process in the system described, it still required additional generations and time at the bench after the point at which mutant isolates were first obtained from the screen. The ease and feasibility of taking

these extra steps of backcrossing, making recombinants, and applying traditional mapping methods to narrow down regions of interest depends heavily on the model system used. Nonetheless, over time, other model organisms with larger—and therefore more expensive—genomes followed suit, using combinations of WGS and traditional mapping to obtain results more efficiently in their forward genetics experiments. Simultaneously, in systems like *A. thaliana*, where the application of WGS to screens was more developed, efforts were underway that tested the limits on the amount of backcrossing necessary to unearth causative mutations (Lindner et al., 2012). This trend continued, and strategies were developed that allowed researchers to be less dependent on high sequencing coverage levels, and to avoid additional generations altogether in their screens. In one such example, an N-ethyl-N-nitrosourea (ENU) mutagenesis screen in mice demonstrated that a combination of filtering for variation caused by ENU, and application of a Lander-Green algorithm for determining identity-by-descent, was sufficient to discover causative mutations for a B cell lymphopenia phenotype (Bull et al., 2013). Eventually, approaches were developed that even removed the necessity for a reference genome sequence, which was extremely beneficial to groups working in non-model organisms (Schneeberger, 2014). The feasibility of conducting WGS on individual genomes obtained from screens, as opposed to pooling many individuals or relying on recombinants, continues to expand. This gives researchers even greater freedom and flexibility in experimental design. As a result, the applicability of WGS to different systems and biological questions in forward genetics continues to expand rapidly.

Despite the undeniable contribution of WGS to forward genetics, its integration into the field has also provided new challenges. As new technologies have continued to emerge, and existing ones have improved in both efficiency and affordability, it has become clear that the work to be done will depend less on WGS itself and more on the filtration and analysis of the overwhelming amount of WGS-generated data (**Figure 1-1b**). The arrival of high-throughput sequencing technologies prompted a demand for solutions to store, process, and analyze this data. Bioinformatics thereby became a major consideration for all labs applying WGS to their forward genetics studies. At first, shorter sequencing reads and the lack of paired-end reads were identified as major challenges in designing and applying bioinformatics pipelines (Pop & Salzberg, 2008). The same bioinformatics pipelines designed to align and assemble sequences from Sanger data, which had been successfully optimized to easily produce reads hundreds of base pairs long, would not be sufficient to complete the same computational tasks with new NGS data. Early NGS users also realized that their data analysis pipelines often needed to be tailored to the types of variants being sought out—whether single nucleotide variants (SNVs), indels, or large structural changes.

The need to customize pipelines for this and other reasons is still prevalent, even with increased read lengths and the availability of paired-end reads. While many of the practical challenges in completing the alignments, quality filtration, and variant calling of WGS data have been addressed, today's researchers are faced with an inundation of choices when it comes to the software, scripts, and pipelines that complete these tasks. One thorough review surveyed

over 200 of these tools and made comparisons in their ability to address several types of questions in human medical genetics (Pabinger et al., 2014). Fortunately, some of the tools discussed—such as SAMtools and the Genome Analysis Tool Kit (GATK)—are also designed for applications in other model systems. This group focused their discussion on the variant calling, annotating, and visualization of variants. Their reasoning for this was twofold. First, the earliest step of quality assessment tends to be built into the sequencing platform itself. Second, the subsequent step of aligning WGS reads had already been reviewed in detail (Li & Homer, 2010; Ruffalo, Laframboise, & Koyutürk, 2011; Yu et al., 2012). In general, findings from those studies indicated that aligners gave similar results for a given WGS data set, but that certain programs were faster or more sensitive to quality thresholds. Other reviews have been published that discuss applications in multiple model systems (Ekblom & Wolf, 2014), and that focus on the challenges inherent in non-model systems (da Fonseca et al., 2016). Finally, databases such as OMICtools contain categorized lists that provide links to different tools for every step of the WGS data analysis process ("Whole-genome sequencing data analysis bioinformatics software tools," 2017). Combining the available information from these resources with findings from representative studies in a given model system can provide a bioinformatic foundation for labs considering the implementation of WGS. In **Table 1-1**, I have provided a summary of examples of the application of WGS to forward genetics studies in multiple model metazoans.

Table 1-1. WGS in forward genetics studies in model metazoans

Study	Sequencing platform used	Data analysis	Results/Notes
(Blumenstiel et al., 2009) EMS mut. screen	Illumina Genome Analyzer I (single-end)	Align: MAQ Variant calling: Directly compare MAQ alignments, filter for EMS-induced changes Candidate filtration/annotation: SIFT, custom scripts	Nonsense mutation in *enc* gene determined causative in fused dorsal appendage phenotype in embryos
(**Laitinen et al., 2010**) Map causative variant in dwarf *A. thaliana*	Illumina Genome Analyzer (paired-end)	Pooled variant calling: SHOREmap, GenomeMapper Candidate filtration/annotation: Dwarf-pool-specific variants assigned by SHORE quality value	Spontaneous frameshift mutation in At1g58440 gene associated with dwarf phenotype
(**Zuryn et al., 2010**) EMS mut. screen in *C. elegans*	Illumina Genome Analyzer II (paired-end)	Align: MAQ Variant calling: MAQ Candidate filtration/annotation: custom scripts	Variant in *egl-5* determined causal in cell re-programming defect. WGS and backcrossing eliminates need for any prior mapping
(**Doitsidou et al., 2010**) *C. elegans* screen for dopaminergic neuron specification deficient mutants	Illumina Genome Analyzer II (single-end)	Align: MAQ Variant calling: MAQ Candidate filtration/annotation: Manual filtration in Excel, data visualization in Adobe Illustrator	Strategy modified from SHOREmap. Premature stop codons in *vab-3* locus identified in deficiencies in dopaminergic neuron specification
(Gerhold et al., 2011) Compensatory growth gene screen	Illumina Genome Analyzer II (combination of single- and paired-end)	Align: MAQ Variant calling: Custom scripts to analyze MAQ coverage and .snp files Candidate filtration/annotation: SIFT, custom scripts	Causative alleles found in *bun*, *RnrL*, and *Top3a*

(Gonzalez et al., 2012) EMS mut. screen	Illumina HiSeq2000 (paired-end reads)	Align: BWA Variant calling: GATK Candidate filtration/annotation: ENSEMBL Variant Effect Predictor (VEP)	Nonsense mutation in *Ect4* determined causative in increased survival of severed peripheral axons
(**Bowen et al., 2012**) Map *D. rerio* SNPs	Illumina HiSeq2000 (single-end)	Align: Novalign Variant calling: SAMtools, BCFtools Candidate filtration/annotation: GATK, custom scripts	First WGS mapping in zebrafish Established SNP database; found causative variants in 2 mutants
(**Tabata et al., 2012**) EMS mut. screen in *A. thaliana*	Applied Biosystems SOLiD 3 Plus System	Align: BioScope 1.3 software Variant calling: BioScope 1.3 Bayesian algorithm Candidate filtration/annotation: Custom scripts	Missense variant in *CTR1* assigned to short root/high-boron requirement mutant phenotype
(**Bull et al., 2013**) Mouse ENU mut. screen	Illumina HiSeq200 (paired-end)	Align: Stampy (BWA settings) Variant calling: Platypus (in-house software) Candidate filtration/annotation: Annovar (Emsembl annotations), Polyphen-2 (trained with HumVar dataset)	WGS identity-by-descent strategy avoids excessive breeding. Missense variant in *Lyn* identified as causal in peripheral B cell lymphopenia
(Haelterman et al., 2014) EMS mut. screen, essential genes	Illumina HiSeq2000 (mostly paired-end, single-end)	Align: BWA, calibrated with GATK Variant calling: Atlas2 variant analysis software Candidate filtration/annotation: custom scripts, remove *D. mel.* Genetic Reference Panel (DGRP) SNVs	Discovery of 274 causative mutations on Chr. X, in 394 EMS-induced mutant strains
(**Lee et al., 2016**) EMS mut. screen	Illumina HiSeq2000 (paired-end)	Align: BWA Variant calling: GATK Candidate filtration/annotation: custom scripts	Mosaic eye assay, WGS, and iPLEX MassARRAY: *Atm*, *Xrp1* mutations affect cell competition
(Sanchez et al., 2017) Map causative *D. rerio* variants, ENU screen, fixed tissue	Illumina HiSeq2500/3000 (paired-end)	Align: Novalign Variant calling: SAMtools Candidate filtration/annotation: snpEFF	Mapped causative mutation for myelination defect in fbxw7 mutants

Abbreviations: MAQ: Mapping and Assembly with Quality | SIFT: Sorts Intolerant From Tolerant (effect of amino acid changes) | BWA: Burrows-Wheeler Aligner | GATK: Genome Analysis Tool Kit | SAMtools: Sequence Alignment Map tools | BCFtools: Binary variant Call Format tools

Comment [FSM3]: This is by far not the only acronym in here that may need explanation

Comment [d4]: Tried to add in the ones that would not be obvious from the text

I have included an over-representation of examples in *D. melanogaster* (shaded rows in **Table 1-1**), as these examples were used to inform the framework that I describe in the following section.

From mutant phenotypes to candidate genes: a start-to-finish framework applied to two *Drosophila* forward genetics studies

For any lab conducting studies in forward genetics, there can be a steep learning curve at each step of the process when integrating WGS. Decisions need to be made about which individuals or pools to sequence, DNA library preparation methods, sequencing platforms, and strategies for data processing, analysis, and organization. *Drosophila melanogaster* is a model organism with decades of history in genetic screens and forward genetics, and is one of many examples of the past, present, and future benefits of WGS in forward genetics. In addition, successful applications within the *Drosophila* system demonstrate that—with the proper data analysis and filtration approaches—WGS can overcome the additional challenge of a constantly evolving genome. This is in contrast to systems such as *S. cerevisiae* and *C. elegans*, where genome evolution can be suspended by freezing down stocks and interesting mutants for long-term storage. A chronological list of representative studies in *D. melanogaster* and other model metazoans and their findings is presented in **Table 1-1**. These studies represent several variations on forward genetics. Despite changes in available technology and software throughout the years, certain programs and platforms become

industry standards for a number of years until improved replacements take hold. For example, while Illumina remains a popular choice of sequencing platform, and many labs depend on GATK for variant handling, popular alignment programs have shifted, from MAQ (Mapping and Assembly with Quality) to BWA (Burrows Wheeler Alignment) to Bowtie/Bowtie2, Novalign, and others.

I have taken the current examples of WGS in *D. melanogaster* forward genetics and analyzed the similarities and differences to forward genetics studies in DNA repair in our own lab. This allowed me to create a WGS and bioinformatics framework that I applied to two studies related to novel roles of genes involved in DNA repair and genomic integrity. In the first study, I generated double mutants lacking two helicases involved in DNA repair, BLM and HELQ. As the genes encoding these proteins are located on opposite arms of the third *D. melanogaster* chromosome, it was feasible to accomplish this using meiotic recombination. I obtained 21 double mutant strains, and only 1 was able to produce viable homozygotes. The other 20 strains displayed a spontaneous synthetic lethality phenotype, with lethality occurring between the third instar larval and pharate adult stage of development. Given the observation that this synthetic lethality was not always present, I proposed that a spontaneous variant in one of the initial mutant stocks (*blm* or *mus301*) had crossed onto the third chromosome. I employed WGS in order to find the third chromosome variant responsible for the lethal phenotype.

The second study was a modification on a classical EMS-mutagenesis screen. The screen was prompted by work from our lab on the translesion

synthesis (TLS) polymerase Rev1. Our lab has previously gathered evidence showing that Rev1 is important to both the TLS and template switching (TS) pathways of DNA Damage Tolerance (DDT) in *D. melanogaster* (Khodaverdian, in preparation). DDT is a pathway in which either TLS or TS is used to bypass DNA lesions during replication in order to prevent more detrimental effects, such as replication fork stalling and collapse. Knocking out the C-terminal domain (CTD) of Rev1 removes its ability to recruit TLS polymerases, and should shunt the DDT pathway toward the use of TS (Khodaverdian et al., in preparation). The role of Rev1 in TS is not well understood. Furthermore, there is still little known about the TS pathway itself, and what other proteins might be involved. In order to address this, our lab conducted an EMS mutagenesis screen in a *Rev1-ΔCTD* mutant background. We screened for extreme sensitivity to methyl-methane sulfonate (MMS) in mutants, indicating potential variants compromising the TS pathway of DDT. These variants could thus result in failure to bypass MMS-induced lesions. Instead of employing traditional mapping methods, I used WGS to discover mutations causing MMS sensitivity in this screen.

For both of these studies, I applied a WGS and bioinformatic data analysis framework that was similar at every step, yet could easily be adjusted in order to account for the differences in the design and goal of each individual study. In addition to deciding on a suitable sequencing platform, library preparation method, and data analysis system, I also implemented best practices for storage and organization of raw WGS data, as well as data created from downstream analysis. Many principles for the latter were adopted or modified from a guide on

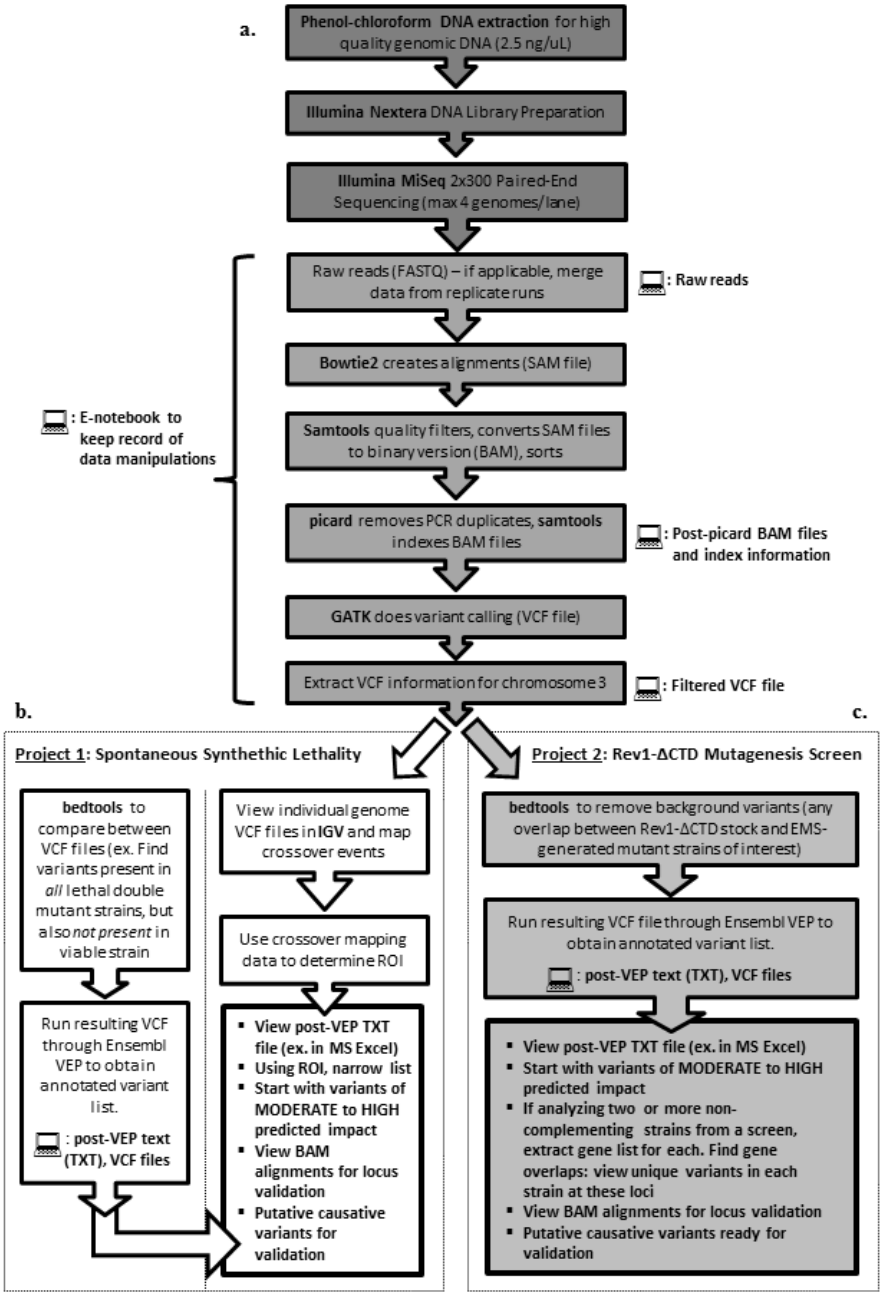


Figure 1-2. A start-to-finish framework for applying WGS to *D. melanogaster* forward genetics. The flowchart shown represents the entire WGS workflow as it applies to two separate forward genetics studies in *D. melanogaster* (see text for further description of these studies). Computer icons represent points at which I recommended saving data (file types indicated) or

keeping records to maintain organizational best practices. (a) Wet-lab and initial bioinformatics workflow, which was similar between both studies. (b) Specific details on downstream workflow customization for Project 1 (*blm* and *mus301* project). (c) Specific details on downstream workflow customization for Project 2 (*Rev1-ΔCTD* EMS mutagenesis screen project).

organizing projects that use bioinformatics/computational biology approaches (Noble, 2009). Although this guide was published at a time when the application of NGS to forward genetics was still a relatively new effort, its principles remain relevant to any research lab planning to do some or all of their own WGS data analysis. The overall framework—including wet lab work, data analysis, file types used, and organizational considerations—is detailed in **Figure 1-2**.

For this framework, I used Illumina's Nextera DNA library preparation kit and the Illumina MiSeq for high-throughput sequencing. To conduct all data analysis in our lab, high processing power was necessary for the various manipulations of raw sequencing data. I accomplished this by using a computing cluster: a connection of multiple local computers with high speed connections accessed remotely (Doughty, 2017). Using the cluster, I was able to take raw sequencing reads, align them to the *Drosophila* reference genome (dm6 build), and generate variant calls for each genome, in the form of VCF files. I could also compare VCF files to one another within the cluster using a program called bedtools, and resultant data were used to generate lists of candidate variants for these two projects. On a per-use basis, I tracked my manipulations of the data within the cluster by keeping an electronic lab notebook (**Figure 1-2a**). Electronic lab notebook entries tracked all activity in a session of cluster use. Keeping these notebook entries allowed recall of previously used commands for running scripts

and working with data sets. Finally, variant annotation was accomplished using the Ensembl Variant Effect Predictor (VEP), which currently services over 80 organisms (McLaren et al., 2016). Results from the VEP can be filtered by predicted impact. I have focused my analyses thus far on variants with moderate to high impact.

The preceding steps represent the generally applicable foundation of the framework (**Figure 1-2a**). However, there are certain places within the framework where I have made adjustments in order to account for the unique requirements of the two different forward genetics projects to which it was applied. Most of these come into play after the point at which an individual genome VCF file is created by GATK. In the *blm mus301* project (Project 1), the genome of each double mutant being sequenced represents a unique third chromosome recombination event. Therefore, the VCF information on each individual's third chromosome represents a pattern of variants corresponding to the crossover events, wherein the homozygous double mutant larval genome can carry one of two possible VCF signatures: one that matches either the *blm* genome or the *mus301* genome. I used the Integrated Genomics Viewer (IGV) to scan the third chromosome for points at which this signature changes (Robinson et al., 2011). This allowed for the successful mapping of recombination events (**Figure 1-3**). This information was used to compare recombination events between the lethal double mutant genomes and the viable double mutant genome, and to seek out a region of interest (ROI) where all lethal genomes share VCF information, yet show no similarity to the

viable genome. I used the ROI to narrow down the list of candidate variants determined by comparing the VCF files with bedtools (**Figure 1-2b**).

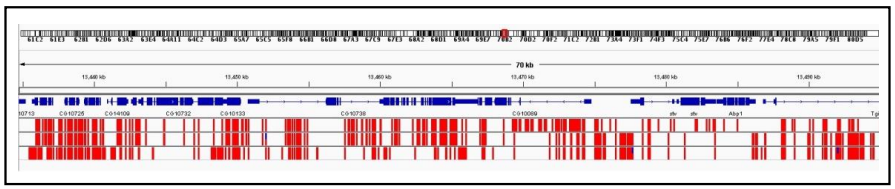


Figure 1-3. Visualizing VCF files in IGV to locate crossovers and map a region of interest. The image above is a representative screenshot from IGV that clearly shows the location of a crossover near the 4th exon of gene CG10089 on the left arm of the third chromosome. All genome sequences were homozygous. The red lines indicate the pattern of variant calls in each of the three genome traces; this information is obtained from the VCF files. The top trace represents the stock *mus301* mutant chromosome. The middle trace represents a single *blm mus301* double mutant. The bottom trace represents the stock *blm* chromosome. The crossover is identified as a location where the double mutant trace switches from resembling the variant call pattern of one mutant stock to resembling the variant call pattern of the other.

In the *Rev1-ΔCTD* project (Project 2), the primary goal was to find the mutations causing MMS sensitivity in mutant strains obtained from an EMS mutagenesis screen. I first applied this to two mutants within a complementation group. However, this dictated that I needed to seek out two unique mutations between these strains, rather than the same mutation in each. Therefore, rather than using bedtools to find the intersection of the VCF information in the two strains, I instead extracted gene lists from each individual strain and then manually scanned this list for overlaps. This allowed the identification of genes with variant calls in both strains, and I could then refer back to the VCF information in order to see if each mutant strain contained unique calls in that

gene (**Figure 1-2c**). In order to visualize and verify variant calls from candidates in both projects, I referred back to alignments and coverage information at the appropriate positions in the genome using IGV for visualization, and this was crucial to filter out low quality and false positive calls.

Applying this framework to these two projects resulted in manageable candidate lists for causative mutations in each. A summary of my process and the results is summarized in **Table 1-2** (Project 1) and **Table 1-3** (Project 2). In **Table 1-2**, I have shown two versions of the same data set. The top row represents the population of variants and genes obtained when considering only the *blm mus301* double mutant strains. The second row represents the same data, but with the *blm* and *mus301* single mutant stocks added to the analysis. Based on the recombination mapping analysis (**Figure 1-2b**, **Figure 1-3**), it is most likely that the causative mutant originated in the *blm* mutant stock. Inclusion of this condition means that the bedtools analysis proceeded as follows: variation in the *mus301* stock and the viable *blm mus301* double mutant is subtracted from the list of all variants present in both the *blm* mutant stock and the population of lethal *blm mus301* double mutants. The latter analysis does provide a slight amount of filtration to the list of candidates. The list of candidates from Project 1 will be analyzed further by viewing the quality of alignments and coverage at the corresponding locations. This will help to filter out false-positive variant calls and provide a more manageable list for validation.

The preliminary analysis from Project 2, which is presented in **Table 1-3,** was based on two mutants from an EMS mutagenesis screen that showed a MMS-sensitivity phenotype, and fell into a single complementation group. That these mutants (148 and 157) belonged to a single complementation group indicated that they contained mutations in the same gene. Both were sequenced, and I found a

Table 1-2. WGS variant data obtained in Project 1 (*blm* and *mus301*)

Data Set Analyzed	Total Variants	Total Variants (HIGH or MODERATE Predicted Impact)	Total HIGH to MOD. Genes	HIGH to MOD. Genes in ROI	HIGH only
Variants in lethals, not in viable	19,954	1166	558	466	35
Variants in lethals and *blm* stock, not in viable or *mus301* stock	16,548	1090	539	394	18

total of 7 genes that had high or moderate predicted impact variants in both 148 and 157. Of these, only one gene—*polybromo*—had unique, high impact variants in each, and also showed high quality alignments and coverage at the relevant locations. This data set therefore indicates *polybromo* as a strong candidate for validation in Project 2. The data obtained from the analyses support the applicability of this framework across two different studies in forward genetics, as well as the framework's ability to provide a manageable list of candidates for causative variants in each.

Table 1-3. WGS variant data obtained in Project 2 (*Rev1-ΔCTD*)

Data Set Analyzed	Total Variants	Total Genes	Genes with HIGH or MODERATE Predicted Impact Variants	Genes HIGH or MOD. Impact, in list for both non-complementing strains
Mutant 148	3,339	1,591	45	7
Mutant 157	3,962	1,666	53	7

Conclusions and future directions

The arrival of NGS technology, and with it the possibility of WGS for any organism, has proved to be an incredible tool in the pursuit of forward genetics questions. It was clear, even in the early days of NGS, that the pre-existing bottleneck imposed on forward genetics by tedious and challenging methods of traditional mapping and variant identification could be lessened, or even eliminated, by this technology. Recent improvements in accessibility and cost have caused a shift in the main challenges that face scientists who hope to implement WGS in their research. The challenge may no longer be the WGS platform itself, but the issues of experimental design, implementation, and most importantly, data handling and analysis (**Figure 1-1b**). In implementing a standard, start-to-finish WGS and bioinformatic data analysis framework, I have maintained consistency of application while still allowing room for customization. To demonstrate this, I presented the preliminary results that were obtained from the application of this framework to two different forward genetics projects from

our lab. Moving forward, the top candidates obtained as outputs from the framework will need to be validated, and this can be done in a variety of ways. For example, in the case of interesting candidates from Project 1, we can look for *D. melanogaster* stocks with inserted transposable elements within the genes of interest. These stocks could then be crossed with the viable *blm mus301* double mutant strain. Progeny from resultant triple mutant recombinant strains could then be screened for larval lethality. Alternatively, CRISPR-Cas9 genome editing could be used to knock out *blm* and *mus301* in a transposable element line for the gene of interest, and resultant triple mutant lines could be similarly screened for larval lethality phenotypes. For any given mutant in Project 2, a candidate causative gene could be expressed as a transgenic wild-type copy in that mutant. Demonstrating subsequent rescue of MMS sensitivity by the transgenic copy of the gene in that mutant would provide significant evidence for validation.

Of course, while I have focused here on forward genetics, similar principles of organization and data analysis can be applied to other studies using WGS. Both forward and reverse genetics studies may employ a pooled sequencing analysis strategy, such as BSA. For these studies, entire software suites such as popoolation2 are available (Kofler, Pandey, & Schlötterer, 2011), which are designed specifically for genomic pools. When designing and planning any project, it is important to find representative examples that outline the application of WGS in the model organism of interest. In doing so, researchers can review specific details pertaining to the best programs to use for data analysis (**Table 1-1**). The list of choices for bioinformatics software and analysis strategies

can be daunting, but this type of preparation will provide valuable guidance, and can ease the transition into WGS workflows. Here, I have described my own framework, which was applied to two studies in forward genetics in our lab, and can now be easily applied to future projects. On a larger scale, it can provide an example of an efficient workflow to labs working in other model systems who wish to use or streamline a WGS workflow.

ACKNOWLEDGEMENTS

I would like to thank Mitch McVey, Justin Blanch, and Tokio Sano for thoughtful comments on this manuscript. I would also like to thank Reazur Rahman for providing the pe_align.sh script, and David LaPointe for his guidance in basic UNIX and working with command line. The research projects described here, to which I applied the WGS and bioinformatics framework, were supported by grant P01GM105473 from the NIH.

REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K A., Howland, T J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., Woodage T., Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., Venter, J. C. (2000). The genome sequence of Drosophila melanogaster. *Science*, *287*(5461), 2185–2195.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815.

Blattner, F. R. (1997). The Complete Genome Sequence of Escherichia coli K-12. *Science*, *277*(5331), 1453–1462.

Blumenstiel, J. P., Noll, A. C., Griffiths, J. A., Perera, A. G., Walton, K. N., Gilliland, W. D., Hawley, R S., Staehling-Hampton, K. (2009). Identification of EMS-induced mutations in Drosophila melanogaster by whole-genome sequencing. *Genetics*, *182*(1), 25–32.

Bowen, M. E., Henke, K., Siegfried, K. R., Warman, M. L., & Harris, M. P. (2012). Efficient mapping and cloning of mutations in zebrafish by low-coverage whole-genome sequencing. *Genetics*, *190*(3), 1017–1024.

Bull, K. R., Rimmer, A. J., Siggs, O. M., Miosge, L. A., Roots, C. M., Enders, A., Bertram, E. M., Crockford, T. L., Whittle, B., Potter, P. K., Simon, M. M., Mallon, A., Brown, S. D. M., Beutler, B., Goodnow, C. C., Lunter, G., Cornall, R. J. (2013). Unlocking the Bottleneck in Forward Genetics Using Whole-Genome Sequencing and Identity by Descent to Isolate Causative Mutations. *PLoS Genetics*, *9*(1), e1003219.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S., Venter, J. C. (1996). Complete Genome Sequence of the Methanogenic Archaeon, Methanococcus jannaschii. *Science*, *273*(5278), 1058–1073.

Cooper, G. M. (2000). The Sequences of Complete Genomes. In *The Cell: A Molecular Approach* (2nd edition). Sunderland, MA: Sinauer Associates.

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., Zepeda-Mendoza, M. L., Campos, P. F., Heller, R., Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, *30*, 3–13.

Doitsidou, M., Poole, R. J., Sarin, S., Bigelow, H., & Hobert, O. (2010). C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. *PLoS ONE*, *5*(11), e15435.

Doughty, S. (2017). High Performance Compute Cluster. Retrieved October 31, 2017, from https://wikis.uit.tufts.edu/confluence/display/TuftsUIT ResearchComputing/High+Performance+Compute+Cluster

Ekblom, R., & Wolf, J. B. W. (2014, November). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*. Wiley-Blackwell.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O., Venter, J. C. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, *269*(5223), 496–512.

Gerhold, A. R., Richter, D. J., Yu, A. S., & Hariharan, I. K. (2011). Identification and characterization of genes required for compensatory growth in Drosophila. *Genetics*, *189*(4), 1309–1326.

Gonzalez, M., Van Booven, D., Hulme, W., Ulloa, R., Lebrigio, R., Osterloh, J., Logan, M., Freeman, M., Zuchner, S. (2012). Whole Genome Sequencing and a New Bioinformatics Platform Allow for Rapid Gene Identification in D. melanogaster EMS Screens. *Biology*, *1*(3), 766–777.

Haelterman, N. A., Jiang, L., Li, Y., Bayat, V., Sandoval, H., Ugur, B., Tan, K. L., Zhang, K., Bei, D., Xiong, B., Charng, W. L., Busby, T., Jawaid, A., David, G., Jaiswal, M., Venken, K. J. T., Yamamoto, S., Chen, R., Bellen, H. J. (2014). Large-scale identification of chemically induced mutations in Drosophila melanogaster. *Genome Research*, *24*(10), 1707–1718.

Huang, S., Romero-Ruiz, M., Castell, O. K., Bayley, H., & Wallace, M. I. (2015). High-throughput optical sensing of nucleic acids in a nanopore array. *Nature Nanotechnology*, *10*(11), 986–991.

Kile, B. T., & Hilton, D. J. (2005). The art and design of genetic screens: mouse. *Nature Reviews Genetics*, *6*(7), 557–567.

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, *27*(24), 3435–3436.

Laitinen, R. A. E., Schneeberger, K., Jelly, N. S., Ossowski, S., & Weigel, D. (2010). Identification of a Spontaneous Frame Shift Mutation in a Nonreference Arabidopsis Accession Using Whole Genome Sequencing. *Plant Physiology*, *153*(2), 652–654.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W.,

Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

Lee, C.-H., Rimesso, G., Reynolds, D. M., Cai, J., & Baker, N. E. (2016). Whole-Genome Sequencing and iPLEX MassARRAY Genotyping Map an EMS-induced Mutation Affecting Cell Competition in Drosophila melanogaster. *G3: Genes|Genomes|Genetics*, *6*(10), 3207–3217.

Li, H., & Homer, N. (2010, September). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, *11*(5), 473–483.

Lindner, H., Raissig, M. T., Sailer, C., Shimosato-Asano, H., Bruggmann, R., & Grossniklaus, U. (2012). SNP-ratio mapping (SRM): Identifying lethal alleles and mutations in complex genetic backgrounds by next-generation sequencing. *Genetics*, *191*(4), 1381–1386.

MacArthur, D. (2010). Ion Torrent bought by Life Technologies. Retrieved October 23, 2017, from https://www.wired.com/2010/08/ion-torrent-bought-by-life-technologies/

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, *24*(3), 133–141.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122.

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46.

Moresco, E. M. Y., Li, X., & Beutler, B. (2013, May). Going forward with genetics: Recent technological advances and forward genetics in mice. *American Journal of Pathology*, *182*(5), 1462–1473.

Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves,

T. A., Green, E. D., Gregory, S., Guigó, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Von Niederhausern, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S. P., Zdobnov, E. M., Zody, M. C., Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562.

National Human Genome Research Institute. (2016). The Cost of Sequencing a Human Genome - National Human Genome Research Institute (NHGRI). Retrieved October 21, 2017, from https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

Noble, W. S. (2009, July 31). A quick guide to organizing computational biology projects. *PLoS Computational Biology*, *5*(7): e1000424.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, *15*(2), 256–278.

Pop, M., & Salzberg, S. L. (2008, March). Bioinformatics challenges of new sequencing technology. *Trends in Genetics*, *24*(3), 142–149.

Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., Gu, Y. (2012). A tale of three next generation

sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*(1), 341.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26.

Ruffalo, M., Laframboise, T., & Koyutürk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, *27*(20), 2790–2796.

Sanchez, N. E., Harty, B. L., O'Reilly-Pol, T., Ackerman, S. D., Herbert, A. L., Holmgren, M., Johnson, S. L., Gray, R. S., Monk, K. R. (2017). Whole Genome Sequencing-Based Mapping and Candidate Identification of Mutations from Fixed Zebrafish Tissue. *G3: Genes|Genomes|Genetics*, g3.300212.2017.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., Smith, M. (1977). Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, *265*(5596), 687–695.

Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., & Petersen, G. B. (1982). Nucleotide sequence of bacteriophage λ DNA. *Journal of Molecular Biology*, *162*, 729–779.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467.

Sboner, A., Mu, X., Greenbaum, D., Auerbach, R. K., & Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome Biology*, *12*(8), 125.

Schneeberger, K. (2014). Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nature Reviews Genetics*, *15*(10), 662–676.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., Jørgensen, J., Weigel, D., Andersen, S. U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods*, *6*(8), 550–551.

Sega, G. A. (1984, September 1). A review of the genetic effects of ethyl methanesulfonate. *Mutation Research - Reviews in Genetic Toxicology*, *134*(2-3), 113–142.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145.

Shendure, J., Mitra, R. D., Varma, C., & Church, G. M. (2004). Advanced sequencing technologies: methods and goals. *Nature Reviews Genetics*, *5*(5), 335–344.

St Johnston, D. (2002). The art and design of genetic screens: drosophila melanogaster. *Nature Reviews Genetics*, *3*(3), 176–188.

Tabata, R., Kamiya, T., Shigenobu, S., Yamaguchi, K., Yamada, M., Hasebe, M., Fujiwara, T., Sawa, S. (2012). Identification of an EMS-induced causal mutation in a gene required for boron-mediated root development by low-coverage genome re-sequencing in Arabidopsis. *Plant Signaling & Behavior*, *8*(1), 5–7.

The *C. elegans* Sequencing Consortium. (1998). Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. *Science*, *282*(5396), 2012–2018.

Van Nimwegen, K. J. M., Van Soest, R. A., Veltman, J. A., Nelen, M. R., Van Der Wilt, G. J., Vissers, L. E. L. M., & Grutters, J. P. C. (2016). Is the 1000 genome as near as we think? A cost analysis of next-generation sequencing. *Clinical Chemistry*, *62*(11), 1458–1464.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K.,

Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351.

Weber, J. L., & Myers, E. W. (1997, May 1). Human whole-genome shotgun sequencing. *Genome Research*, *7*(5), 401–409.

Wetterstrand, K. (2016). DNA Sequencing Costs: Data - National Human Genome Research Institute (NHGRI). Retrieved October 21, 2017, from https://www.genome.gov/27541954/dna-sequencing-costs-data/

Whole-genome sequencing data analysis bioinformatics software tools. (2017). Retrieved October 30, 2017, from https://omictools.com/whole-genome-resequencing-category

Williams, N. White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L., Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H., Fraser, C. M., Smith, H. O., Woese, C. R., Venter, J. C. (1996).

Genome Projects: Yeast Genome Sequence Ferments New Research. *Science*, *272*(5261), 481–480.

Yu, X., Guda, K., Willis, J., Veigl, M., Wang, Z., Markowitz, S., Adams, M. D., Sun, S. (2012). How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Mining*, *5*(1), 6.

Zuryn, S., Le Gras, S., Jamet, K., & Jarriault, S. (2010). A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics*, *186*(1), 427–430.

Chapter 2

Using Whole Genome Sequencing (WGS) to investigate a partial synthetic lethality in *blm mus301* mutants in *Drosophila melanogaster*

Contributions to this work:

Creation of blm^{N1} and $mus301^{288A}$ alleles: Mitch McVey

Designed and provided script for generating variant call format files from raw sequencing data: Reazur Rahman

Illumina MiSeq runs: Alexander Ferrazolli

ABSTRACT

DNA damage is a constant obstacle in a cell's struggle to maintain genomic integrity. In response to this threat, multiple repair mechanisms have evolved to avoid the accumulation of damage-induced lesions, which can ultimately lead to large chromosomal defects and rearrangements, aberrant growth, or cell death. Among the classes of proteins that participate in DNA repair, helicases are of particular interest because they are essential not only to many different repair pathways, but also to normal replication and meiotic recombination. In *Drosophila melanogaster*, conserved 3' to 5' DNA helicases Bloom (DmBlm) and HelQ (DmHelQ) both have important roles in the maintenance of genomic stability, specifically in DNA double-strand break repair (DSBR). The genes encoding DmBlm (*blm*) and DmHelQ (*mus301/spn-C*) are both located on the third chromosome, and mutants in either gene result in defects in both DSBR and development. However, the roles and interactions of DmHelQ, in particular, remain less well defined. I used meiotic recombination to generate *blm mus301* double mutant strains in order to investigate genetic interactions of these helicases. I discovered a synthetic, larval-stage homozygous lethality that was present in most, but not all, *blm mus301* double mutant strains. I proposed that a third mutation had caused the larval lethality phenotype, and applied whole genome sequencing (WGS) to conduct recombination mapping and to find candidates for a causative variant. Based on data obtained from our variant call format (VCF) file mapping data, it appears that the causative mutation originated on the *blm* chromosome, and that lethality is only apparent in the absence of

mus301 and the presence of this new variant. A bioinformatics and data analysis pipeline was developed to create a candidate list with several candidates for the causative mutation. Considering the current gap in knowledge on HelQ in *D. melanogaster*, these findings have the potential to shed light on the function of this important protein.

INTRODUCTION

Helicases are best known as the proteins that anneal and unwind DNA strands. Perhaps the most well-known helicases are the major replicative MCM-family helicases, which travel ahead of replication forks in order to open up the double-stranded DNA for access by the molecular replicative machinery. These elegantly designed molecular motors use ATP hydrolysis to unwind the double-helical DNA molecule, and have also been shown in mice and yeast to anneal DNA as part of the replication stress response (Sheu, Kinney, Lengronne, Pasero, & Stillman, 2014; You & Masai, 2017). However, beyond the standard definition of helicases lies a vast complexity, not only in the variety of different helicases at work in most organisms, but also in each of their preferred substrates or structure specificity, their strand polarity, and their specific functions. In addition to their role in replication, helicases are also essential to processes that preserve genomic integrity. Many of these processes take the form of DNA repair pathways, and include the resolution of complex endogenous structures, such the Holliday junctions (HJs) created by DNA recombination, and G-quadruplex (G4) DNA

(Mendoza, Bourdoncle, Boulé, Brosh, & Mergny, 2016). Not surprisingly, the central involvement of helicases in pathways of DNA repair implicates them in a variety of cancers. Helicases pose somewhat of a double-edged sword in the context of cancer. DNA lesions that might persist in the absence of helicases can have carcinogenic potential. However, the rapid divisions of cancer cells can also depend on the upregulation of helicases in order to maintain stability and resist the effects of chemotherapeutic agents (Brosh, 2013). Overall, the sum of helicase involvement in cancer makes this class of proteins incredibly important as a subject of research and as a potential target in drug development.

The many roles of BLM in DNA repair and genomic integrity

The Bloom's Syndrome helicase, BLM, is an incredible example of a helicase that exercises its potential in many ways, and within many pathways. In humans, mutations in *BLM* are the cause of Bloom's syndrome, which is characterized by short stature, infertility, immunodeficiency, and a predisposition to multiple cancers, as well as an increased likelihood of early cancer onset (Brosh, 2013; de Renty & Ellis, 2017). BLM is a 3' to 5' ATP-dependent helicase (Nathan A Ellis et al., 1995) which belongs to the subfamily of Superfamily II (SFII) helicases named for the *Escherichia coli* homologous recombination repair (HRR) helicase, RecQ (Nakayama et al., 1984). Most eukaryotes possess more than one RecQ ortholog. Among these orthologs, BLM, in particular, is highly conserved, with homologs in model eukaryotes from yeast to humans. The level

of conservation is strongest around the core SFII helicase domain, indicating the importance of BLM's ATP-dependent helicase activity to many of its functions in the cell. However, certain functions of BLM, such as its impact on gene conversion tract length in non-crossover HR pathways, seem to be independent of its helicase activity (Ertl et al., 2017).

Various *in vitro* biochemical analyses provided important foundations for understanding BLM's role in processes of genetic maintenance. It was demonstrated that BLM could actively resolve single and double Holliday junctions (dHJs) (Wu & Hickson, 2003), regressed replication forks (MacHwe, Karale, Xu, Liu, & Orren, 2011), and G4 structures (Mohaghegh, Karow, Brosh, Bohr, & Hickson, 2001). BLM has also been shown to have strand annealing capability (Cheok, Wu, Garcia, Janscak, & Hickson, 2005), to participate in HJ branch migration (Karow, Constantinou, Li, West, & Hickson, 2000), and to disrupt displacement loops (D-loops) (Van Brabant et al., 2000). Early *in vivo* and *in vitro* evidence showed BLM interacting with the recombinase Rad51 and co-localizing to sites of irradiation induced damage in human cells (Wu, Davies, Levitt, & Hickson, 2001). The same study also demonstrated an interaction between Rad51 and the yeast BLM homolog, Sgs1. Another set of studies showed BLM being recruited to the sites of stalled replication forks, along with the p53 tumor suppressor and Rad51, following hydroxyurea treatment (Sengupta et al., 2003). This group also showed that p53 localization to these sites and its interaction with Rad51 was dependent on BLM. Finally, they showed that damage sensor protein 53BP1 interacted with BLM at these sites in a process dependent

on the Chk-1 kinase, a mediator which signals replication stress (Sengupta et al., 2004). These studies laid out a clear role for BLM in homology-directed repair (HDR) during replication. A set of *in vivo* experiments in mice in 2010 demonstrated increased chiasmata formation in conditional knock-outs of *Blm*, and was the first evidence supporting a role for BLM in mammalian meiosis (Holloway, Morelli, Borst, & Cohen, 2010). These findings were in general agreement with prior studies on *sgs1* in yeast (Ira, Malkova, Liberi, Foiani, & Haber, 2003). Taken together, all of this evidence assigned BLM to pathways in DSBR, Synthesis Dependent Strand Annealing (SDSA), and Homology Directed Repair of regressed four-way junctions resulting from replication fork stall or collapse (reviewed in Brosh, 2013; Payne & Hickson, 2009). Support for these assignments was lent by descriptions of phenotypes in *BLM* mutants, including increased sensitivity to DNA damaging agents and replicative stress; increased genomic instability manifesting in the form of sister chromatid exchanges, translocations, and large deletions; increased tumorigenesis; and deficiencies in DNA gap repair assays (Melissa D Adams, McVey, & Sekelsky, 2003; Davalos & Campisi, 2003; Davalos, Kaminker, Hansen, & Campisi, 2004; N A Ellis, Proytcheva, Sanz, Ye, & German, 1999; Gruber, 2002; McVey, Andersen, Broze, & Sekelsky, 2007; Suzuki, Yasu, & Honma, 2016; Ui et al., 2001). Finally, studies of human BLM demonstrated its ability to promote end resection shortly after the occurrence of a meiotic DSB, underlining the additional ability of BLM to act much earlier on in a repair pathway and to potentially play a role in repair

pathway choice (Grabarz et al., 2013; Nimonkar et al., 2011; Nimonkar, Ozsoy, Genschel, Modrich, & Kowalczykowski, 2008).

The incredible collection of studies on BLM has changed researchers' characterization of the helicase since its discovery. BLM has often been primarily referred to as an anti-crossover helicase: acting to avoid the crossover products of recombination by promoting SDSA over DSBR, dissolving dHJs in meiotic recombination, and also by preventing promiscuous mitotic crossovers. While this still remains an important aspect of BLM's job description, it has also become clear that BLM can act both early and late in these pathways, with many different proteins, and at various points in the cell cycle—in either a pro- or anti-recombinational matter, depending on the cellular context (Davalos et al., 2004; Maréchal & Zou, 2013).

Roles of HELQ in DNA repair and genomic integrity

In contrast to BLM, HELQ/HEL308 seems to be more specialized in its function, acting primarily alongside the FANC-family proteins in interstrand crosslink (ICL) repair, and with Rad51 paralogs during HDR at damaged replication forks. Similarly to BLM, mutations in HELQ have been implicated in certain cancers—such as breast and ovarian cancer—although it appears that this association only holds true in certain human populations (Hamdi et al., 2016; Han, Zhao, & Li, 2016; Pelttari et al., 2016; Rosales-Nieves & González-Reyes,

2014). HELQ deficiencies may also signify risk for early-onset menopause (Stolk et al., 2012). HELQ is an ATP-dependent SFII DNA helicase with 3' to 5' strand unwinding activity. Its level of conservation nearly matches that of BLM, with homologs in model metazoans from *C. elegans* through humans, and some archaea (Gramates et al., 2017; Woodman, Brammer, & Bolt, 2011). HELQ was initially discovered as HEL308, and was identified by its similarity to the helicase-like domain of the Mus308/PolQ/Polθ protein in *D. melanogaster* (Marini & Wood, 2002), which is important for ICL repair and repair of DSBs by microhomology-mediated end-joining (MMEJ) (Beagan et al., 2017; Chan, Yu, & McVey, 2010). The 2002 Marini & Wood study also demonstrated the preference of HELQ to single stranded substrates, and that its activity could be stimulated by DNA single-stranded binding proteins such as RPA. In contrast to Mus308's mechanism of ICL repair by MMEJ, further *in vitro* and *in vitro* evidence has shown that HELQ likely accomplishes ICL repair within a recombination-based pathway. HELQ was shown to localize, along with Rad51, to sites of active replication following treatment with camptothecin, a topoisomerase inhibitor which causes replication stalling and collapse (Tafel, Wu, & McHugh, 2011). In *C. elegans*, a meiotic synthetic lethal interaction was discovered between the Rad51-paralog *rfs-1* and *helq-1* (Ward et al., 2010). Rad51-paralogs are important for Rad51 filament assembly and disassembly during recombination-mediated repair. As the only known Rad51-paralog in *C. elegans*, *rfs-1* has been shown to be essential to recruiting Rad51 to blocked replication forks (Ward, Barber, Petalcorin, Yanowitz, & Boulton, 2007). In the 2010 Ward et al. study, because

lethality only occurred when both *rfs-1* and *heq-1* were compromised, this argued for a role for *helq-1* in Rad51 disassembly as well, and in fact, this group went on to demonstrate that both HELQ-1 and RFS-1 could effectively displace Rad51 from dsDNA. RFS-1 was also shown to interact with RTEL-1, which can disrupt D-loop structures in a mechanism similar to that of BLM (Adelman & Boulton, 2010), lending further evidence to the meiotic and mitotic pathways of repair and genomic stability in which BLM and HELQ may both be involved.

Blm and HelQ in *Drosophila melanogaster*

In *D. melanogaster*, both Blm and HelQ were originally isolated from a large mutagenesis screen, and named for their *mus* (mutagen sensitive) phenotypes as *mus309* and *mus301*, respectively (Boyd et al., 1981). Mutants isolated from this screen for *blm*/*mus309* and *mus301* were each sensitive to the alkylating agent methyl-methane sulfonate (MMS), and to the nitrogen mustard, bis(2-chloroethyl)methylamine (HN2). Although the *D. melanogaster* NHEJ protein Ku70 was originally assigned to the *mus309* locus, further analysis indicated that DmBlm was the protein product of *mus309* (Kusano, 2001). Early characterization of DmBlm confirmed its role in promoting SDSA, and more specifically, in the post strand-invasion disruption of D-loops (Adams, McVey, & Sekelsky, 2003; McVey, Larocque, Adams, & Sekelsky, 2004). Other work showed that *blm* mutants were unable to resolve dHJ, and also that they were synthetically lethal with mutations in the endonuclease *mus81*, which resolves

Comment [FSM5]: Odd formatting of this reference

Comment [BS6]: The first ref has a first name and middle initial in there; I removed

various recombination intermediates. This synthetic lethality was not rescued by loss of Rad51, indicating DmBlm's role in replication-mediated pathways (Johnson-Schlitz & Engels, 2006; Trowbridge, McKim, Brill, & Sekelsky, 2007). By the time this work had been published on DmBlm, several studies had already shown that the *mus301* (*spn-C*) locus, which would later be recognized as DmHelQ, was involved in proper oocyte patterning and development (Ghabrial & Schüpbach, 1999; González-Reyes, Elliott, & St Johnston, 1997). The connection between *mus301* and DmHelQ was established in a study that demonstrated the similar of the *mus301* locus to that of human and mouse HEL308. In the same study, increased damage in *D. melanogaster* oocytes—assayed by the accumulation of phosphorylated γH2AV—demonstrated the importance of *mus301* in meiotic progression and recombinational repair. Beyond the results described by McCaffrey et al. 2006, roles of HelQ in *D. melanogaster* have not been well-studied.

In the work described here, I used characterized null alleles for *blm* and *mus301*. Both alleles used have large deletions that remove or compromise the genes' conserved helicase domain (**Figure 2-1**). When the blm^{N1} allele that we used was first made by imprecise P-element excision and characterized, the authors discovered by reverse-transcriptase PCR that truncated transcript was present in this mutant (McVey et al., 2007). However, genetic data from the same study established that blm^{N1} was genetically null. Evidence for this included extremely low hatch rates and hypersensitivity to ionizing radiation (nearly 0% relative survival at 2,500 rads) in blm^{N1} homozygotes. In addition, helicase-dead

blm^{N1}/blm^{D2} heteroallelic mutants displayed severe defects in gap repair by SDSA—including large flanking deletions—and increased mitotic crossovers. The *mus301* allele we used was $mus301^{288A}$, which was also created by P-element excision, but in a different genetic background then that of blm^{N1}. We have shown previously that $mus301^{288A}$ mutants have increased sensitivity to HN2 and to Topetecan (Thomas, unpublished data). HN2 is a DNA alkylating agent, which can induce ICLs (Kohn, Spears, & Doty, 1966; Siede, 2014). Topetecan is a Topoisomerase I inhibitor which traps the enzyme in the cleavage complex, and this can lead to replication fork collapse and DSBs. The sensitivity of *mus301* mutants to these DNA damaging agents was consistent with prior findings on HELQ in *C.* elegans, mice, and human cell lines (Adelman et al., 2013; Muzzini, Plevani, Boulton, Cassata, & Marini, 2008; Takata, Reh, Tomida, Person, & Wood, 2013). We also demonstrated that the $mus301^{288A}$ mutant behaved similarly to blm^{N1} mutant in terms of its defects in gap repair, although it did not show the large flanking deletions characteristic of the blm^{N1} mutant (Thomas, unpublished data). The same study from our lab revealed that combining these genes in a $mus301^{288A}$ blm^{N1} mutant resulted in rescue of the large flanking deletion phenotype, implying that DmHelQ likely acts upstream of DmBlm in SDSA. Additionally, we had previously observed in another study that we could obtain both viable and inviable $mus301^{288A}$ blm^{N1} double mutant strains (Carrie Hui, unpublished data). Together, these results prompted us to pursue the genetic basis of $mus301^{288A}$ blm^{N1} mutant lethality.

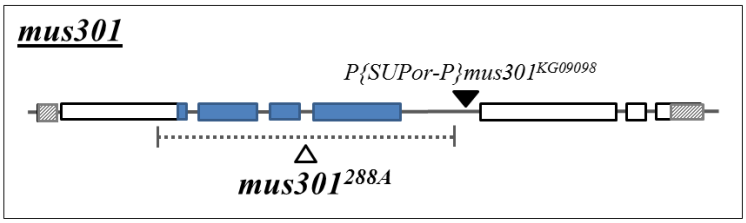
D. melanogaster provides an excellent model in which to study novel interactions in DNA repair, as techniques in both forward and reverse genetics and genetic manipulation are robust and well-established in this system. Additionally, the availability of whole genome data (M D Adams et al., 2000) and WGS technology has improved on these well-established techniques, while also opening up new avenues of study. Here, I apply these tools to the helicases DmBlm and DmHelQ in order to gain further understanding of their roles in repairing and safeguarding of the genome.

RESULTS

Synthetic larval lethality in *blm mus301* double mutants

In order to study interactions between the *blm* and *mus301* helicases in repair and genomic integrity, I created *blm mus301* double mutants. I started with the alleles described previously. The blm^{N1} allele has a large deletion of 2,480 bp, which eliminates the first two conserved motifs in the helicase core domain. The $mus301^{288A}$ allele has a large deletion of 2,069 bp, which completely removes the conserved helicase domain. Despite being genetically null mutants in terms of SDSA/DSBR, as discussed previously, both the blm^{N1} and the $mus301^{288A}$ stocks

a.



b.

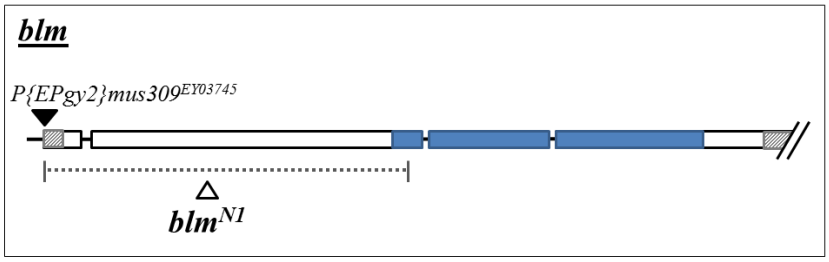


Figure 2-1. **Mutant alleles used for *blm* and *mus301*.** Alleles used for creation of *blm mus301* double mutants are shown. Regions containing the conserved helicase motifs are shown in the blue shaded regions. Gray hatched regions represent 5' and 3' untranslated regions (UTRs). (a) *The mus301^{288A}* allele was created by imprecise excision of the *P{SUPor-P}mus301^{KG0908}* transposable element in the fourth intron. Excision resulted in an upstream deletion of 2,069 bp. This deletion leaves the promoter and transcription and translation start sites intact, but removes part of the first exon, as well as the entire second, third and fourth exons. This deletion effectively removes the entire conserved helicase region. (b) Allele diagram modified from McVey, Andersen, Broze, & Sekelsky, 2007. The *blm^{N1}* allele was created by imprecise excision of the *P{EPgy2}mus309^{EY03745}* transposable element in the 5' UTR of *blm*. Excision resulted in a deletion of 2,480 bp. An intact promoter is preserved, but the first two conserved helicase motifs are removed.

are able to produce viable homozygotes. Because these genes are both located on the third chromosome (**Figure 2-2)**, and are also located at a large genetic distance (28.1 Mbp) apart from one another, the natural process of meiotic recombination during female gametogenesis allows the generation of third chromosome recombination events in the germline of *blm^{N1}/mus301^{288A}* transheterozygous females. After crossing these females to balancer males, I

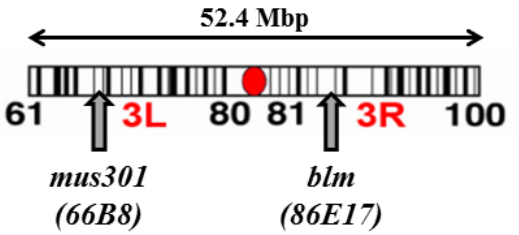


Figure 2-2. Cytogenic map diagram showing approximate locations of *mus301* and *blm* on *D. melanogaster* chromosome 3. The b*lm* and *mus301*genes are located 28.1 Mbp apart from one another. Cytogenic map locations are provided in parentheses below gene names. Modified from Roote & Prokop, 2013.

collected 50 individual male progeny and crossed each male to three balancer females. After 5-6 days of mating, single males from these crosses were sacrificed and genotyped by PCR for both the blm^{N1} and $mus301^{288A}$ alleles using deletion-flanking primers. Of the 50 males genotyped, I obtained 13 indicated successful recombination events (i.e., single males for which both deletion alleles were confirmed). Progeny from the crosses of these 13 individual males were collected, and double mutant siblings were interbred to establish double mutant stocks. Of these 13 stocks, 12 showed a homozygous synthetic lethality phenotype, but 1 was able to produce viable homozygotes at Mendelian ratios (**Table 2-1**). Out of the 13 blm^{N1} $mus301^{288A}$ stocks, 4 were lost due to bacterial contamination (shaded rows in **Table 2-1**), and could not be analyzed further.

Table 2-1. Characterization of double mutant strains as homozygous lethal or viable.

Double mutant strain	Percent homozygous
bm-6 (lethal)	3.85 (n=156)
bm-13	N/A
bm-14 (viable)	24.51 (n=306)
bm-16 (lethal)	0 (n=83)
bm-17(lethal)	1.98 (n=202)
bm-18 (lethal/viable?)	9.63 (n=218)
bm-22 (lethal)	0 (n=110)
bm-33 (lethal)	1.39 (n=144)
bm-37 (lethal)	4.92 (n=244)
bm-41 (lethal)	0.51 (n=198)
bm-43 (lethal)	0 (n=215)
bm-47 (lethal)	1.49 (n=201)
bm-48 (lethal)	1.70 (n=353)

Naming format for double mutant strains is as follows: ***bm-6***, (***b****lm-****m****us301* - followed by the number (***6***) of the cross, out of the 50 total single male crosses). Shaded rows represent stocks lost to bacterial contamination, and were therefore not sequenced or subjected to VCF mapping. n values represent the total number of flies counted.

The remaining collection of strains—8 homozygous lethal and 1 homozygous viable—was subjected to developmental observation on grape agar plates, in order to determine the timing of the synthetic lethality. All 9 double mutant stocks were generated over a GFP-containing balancer chromosome, and homozygotes were identified by a lack of GFP expression. Observations of GFP- larvae on grape agar plates revealed that, in all 8 homozygous lethal blm^{N1} $mus301^{288A}$ strains, the synthetic lethality was consistently occurring between late 3rd instar larval and pharate adult stage (**Figure 2-3a**). These double mutants were created by recombination events on the third chromosome, and so I hypothesized that

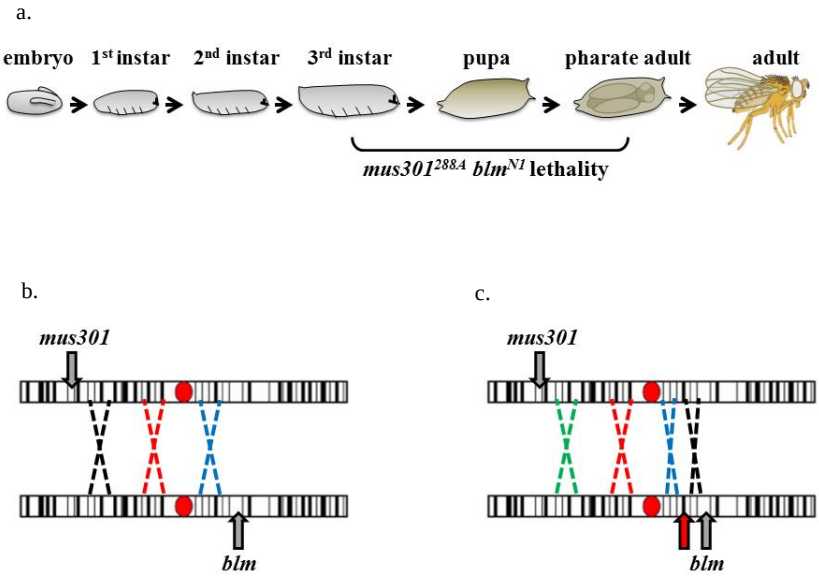


Figure 2-3. Synthetic larval lethality in *blm^{N1} mus301^{288A}* double mutants and potential causative third mutation crossover events. (a) The stages of *D. melanogaster* development from embryo to adult are shown. The bracket represents the range in development at which lethality occurs in all 8 homozygous lethal double mutants. Development stages image modified from Andersen, Kuo, Savukoski, Brodsky, & Sekelsky, 2011. (b) Representation of potential recombination events of the third chromosome, occurring in the female germline in *blm^{N1}/mus301^{288A}* transheterozygotes, assuming a causative variant for larval lethality originates in the *mus301^{288A}* stock. The gray arrows represent positions of *mus301* and *blm* on the third chromosome, and a red arrow represents a potential third mutation. Red, green, and blue dotted X's represent potential, individual crossover events which would lead to the inclusion of the third mutation in one of the meiotic products. A black X represents an event that would not include the third mutation. (c) Representation of potential recombination events of the third chromosome, occurring in the female germline in *blm^{N1}/mus301^{288A}* transheterozygotes, assuming a causative variant for larval lethality originates in the *blm^{N1}* stock. Labeling is the same as in (b). Third chromosome diagrams are modified from **Figure 2**. For simplicity, only single crossover events are shown here.

the synthetic larval lethality had been caused by a third mutation which had crossed over with either *mus301^{288A}* or *blm^{N1}* (**Figure 2-3b, c**). The production of homozygous adults at Mendelian ratios in the single viable *blm^{N1} mus301^{288A}* strain (*bm-14* in **Table 2-1**) provided evidence for the involvement of a single causal gene in the synthetic lethality phenotype. Finally, the crosses and analyses

described above were repeated in an attempt to obtain a higher number of recombination events for more powerful mapping and downstream data analysis. Mapping of recombination events allows the generation of a region of interest for causative variant candidates (see below); therefore, having a greater number of recombinants became important to narrowing the region of interest. The second round consisted of 270 single male crosses, and after genotyping, 20 new double mutants were identified. Due to bacterial contamination issues, 8 of these stocks were lost. The remaining 12 stocks were all homozygous lethal strains. One of these, strain *bm2-161*, did produce a small number of homozygotes (8% homozygotes, n=65 total flies). However, because homozygotes were not produced at Mendelian ratios as in the *bm-14* viable strain, the *bm-2-161* strain was categorized as lethal. This categorization was also supported by the information obtained through recombination mapping (see below).

Comment [BS7]: Added in

Comment [BS8]: I figured this explanation might not be as relevant here anymore, so I am tentatively removing it

Using WGS to find candidates for a causative variant

To work toward identification of a causative mutation, I first subjected all 9 double mutants from the round one crosses, as well as the original single mutant stocks, to WGS. Before WGS, genomic DNA was isolated from homozygous flies (for blm^{N1} and $mus301^{288A}$ single mutant stocks) or larvae (for all blm^{N1} $mus301^{288A}$ double mutants) using a phenol-chloroform extraction. The subsequent DNA library preparation relies on clean, high-quality genomic DNA, and so we validated genomic DNA by gel electrophoresis, Nanodrop, and finally

by DNA, RNA, and Protein Qubit. Any genomic preps with contaminating RNA were treated with RNaseA and column purified. I then proceeded to use the Illumina Nextera kit to create DNA libraries.

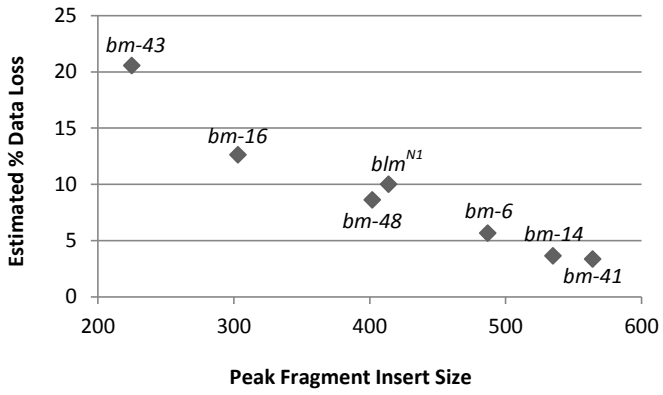


Figure 2-4. Estimated percentage data loss as a function of peak fragment insert size. Each point represents data from an individual genome, which was sequenced via 2x300 paired-end sequencing runs on the Illumina MiSeq. The points shown comprise a representative set of data from a total of 2 sequencing runs, but do not include all genomes used for downstream analysis.

I checked the library fragment size distribution using an Advanced Analytics Fragment Analyzer. Lower peak fragment sizes tended to correlate with higher levels of data loss during the subsequent sequencing reaction (**Figure 2-4**). This is because, in a 2x300 paired end run designed produce 300 bp read lengths, library fragment inserts shorter than 300 bp will be read through to the point of the adapter sequences. Detection of adapter sequences leads to read trimming, and this can result in a certain percentage of data loss for that fragment. Based on

downstream analysis, the amount of data loss experienced in these genomes was a non-issue in comparison to the amount of available useable data. However, for the sample with the greatest data loss, *bm-43* (**Figure 2-4**), a new genomic prep and DNA library was made, and this sample was submitted on another WGS run. Illumina MiSeq was used to sequence all genomic preps using 2x300 paired-end sequencing runs. For the for blm^{N1} and $mus301^{288A}$ stocks, two separate genomic preps, library preps, and sequencing runs were done for each. Raw read data from these runs was concatenated to build coverage. Most blm^{N1} $mus301^{288A}$ double mutants were sequenced only once, but for any double mutants run on an Illumina QC chip, or prepared and run on a second full WGS run, these raw data were also concatenated in a similar manner to take advantage of any added coverage.

In order to find causative variation, I sought out variants that segregated at 100% with one phenotype (homozygous larval lethality), and simultaneously at 0% with the other (homozygous viable). The presence of variants in the various double mutant strains which either overlapped with the single mutant stocks (background), or differed between the various lethals (stock divergence), was not a concern, as these could easily be filtered out computationally. Because ultra-high coverage was not a necessity for these analyses, I did not optimize for the maximum coverage level that would be obtained by running only one genome per lane, and instead opted to run up to four indexed genomes per lane on the MiSeq flowcell.

Applying bioinformatics-based data analysis framework to WGS data

I applied a bioinformatics pipeline (pe_align.sh) to take the raw sequencing output from WGS and generate variant calls for each genome in the form of VCF files. Variant calls were made by comparison to the most recent build of the *Drosophila melanogaster* genome: dm6 (Dos Santos et al., 2015). First, to accomplish recombination mapping and determine a region of interest (ROI) for a causative variant, I took a novel approach in mapping by viewing VCF files directly using Integrated Genomics Viewer (IGV). IGV displays variant calls as red vertical lines that make up a trace for each genome, and this trace is aligned to the reference genome (**Figure 2-5a**). For each sequenced double mutant genome, I viewed its VCF pattern alongside the patterns in the blm^{N1} and $mus301^{288A}$ stock genomes. I used a 70 kb sliding window to scan the entire third chromosome. At any given location, the pattern of variant calls in the double mutant third chromosome was characterized as blm^{N1}-like or $mus301^{288A}$-like. By looking for areas where variant call patterns shifted from blm^{N1}-like to $mus301^{288A}$-like, or vice versa, I could identify regions where recombination had taken place. In most instances, the location of the recombination event could be easily assigned to a specific cytogenic map location. However, there were cases where the variant call pattern was more similar between the three traces around the area of the recombination event, such that the exact location could only be estimated in a broader range.

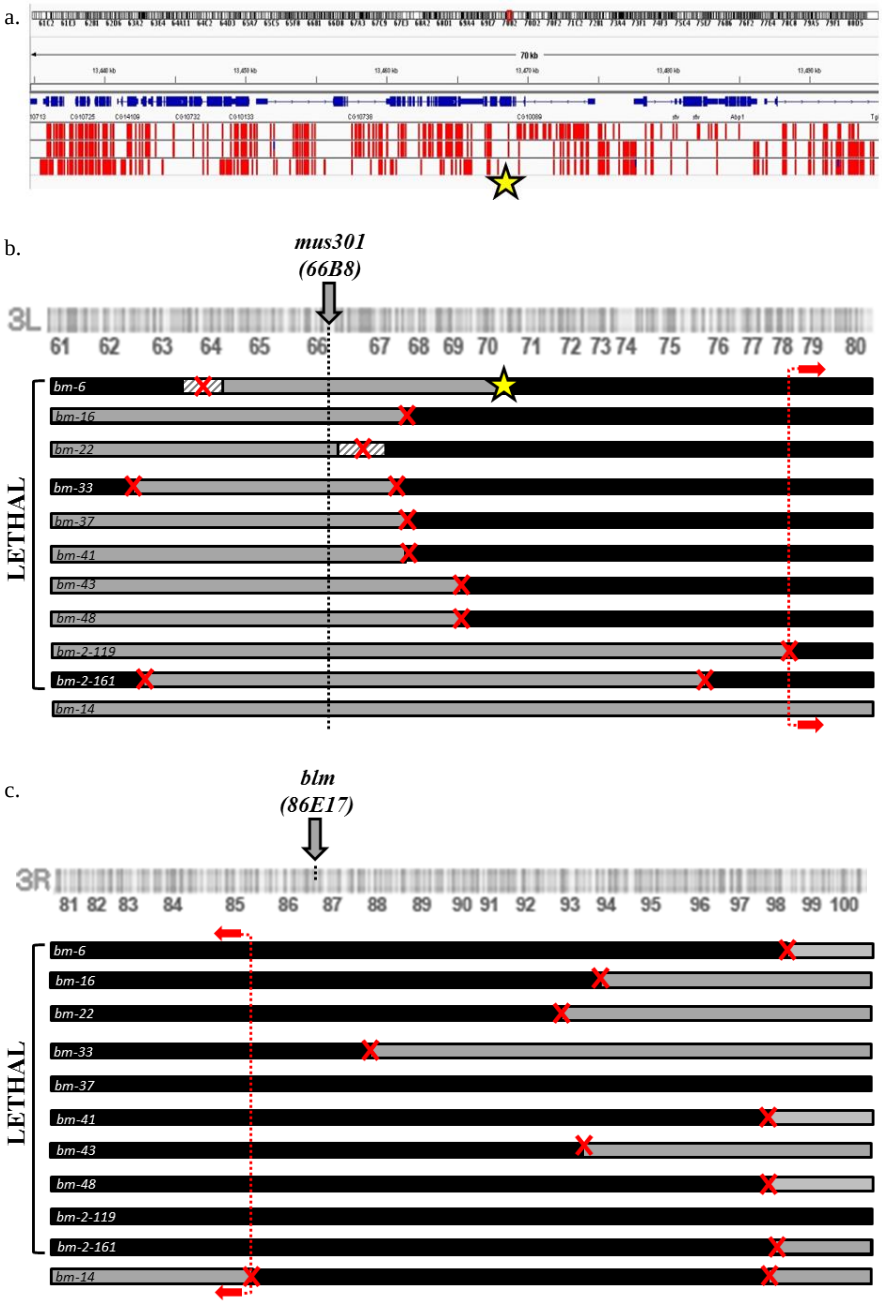


Figure 2-5. **Identifying crossover events and determining an ROI using VCF recombination mapping (VRM)**. VCF files were viewed in IGV, and crossover events were mapped to approximate locations along the third chromosome to determine an ROI for a causative variant. (a) Example screenshot from IGV that clearly shows the location of a crossover near the 4th exon of

gene CG10089 on the left arm of the third chromosome, at cytogenic map location 70B2. Red lines indicate variant calls in each of the three genome traces. The top trace represents the stock *mus301*^{288A} mutant chromosome. The middle trace represents the *bm-6* double mutant. The bottom trace represents the stock *blm*^{N1} chromosome. The location of the crossover, which was one of the ROI defining events, is labeled with a yellow star. (b) Recombination maps for the left arm of chromosome three. Each bar represents a double mutant genome. Black sections have a *blm*^{N1}-like VCF pattern and gray sections have a *mus301*^{288A}-like VCF pattern. Hatched gray and white sections are areas of uncertainty (described in text). Crossover events are represented as red X's. The ROI-defining crossover, also shown above in (a), is labeled again with a yellow star. (c) Recombination maps for the left arm of chromosome three. Labeling is the same as in (b). In (b) and (c), red dotted brackets with arrows indicate ROI boundaries. 3L and 3R cytogenic diagrams modified from the DGRC website (http://fruitfly.jp/flystock/index_e.html)

The collective results of the VCF recombination mapping (VRM) strategy is shown for the left (**Figure 2-5b**) and right (**Figure 2-5c**) arms of chromosome three. Fortunately, the crossovers defining the ROI were among the more clearly defined recombination events. Analysis of the round one strains resulted in an ROI between cytogenetic map locations 70B2 (crossover on 3L in *bm-6* lethal strain) and 85B7 (crossover on 3R in *blm-14* viable strain), which corresponded to a distance of approximately 20 Mbp.

I applied bedtools software to compare the VCF files in a way that would generate lists of candidate variants. Each individual genome VCF file contained all third chromosome variants called by comparison to the dm6 *D. melanogaster* genome sequence. The bedtools software allowed VCF datasets to be intersected and subtracted from each other. I first intersected the variants calls for all of 8 lethal double mutants. From this intersection, I then subtracted the variants present in the *bm-14* viable double mutant. The VRM analysis determined an ROI region which indicated that a causative variant had likely originated on the *blm*^{N1} chromosome (**Figure 2-5b, c**). To account for this finding, I also conducted a

separate stock-inclusive bedtools analysis, which intersected all lethal double mutants with the variants in the *blm*^{N1}stock. From this intersection, I subtracted the variants in the *bm-14* viable double mutant *and* those in the *mus301*^{288A} stock. The resultant post-bedtools VCF files for both analyses were annotated using the Ebsembl Variant Effect Predictor (VEP). For each data set analyzed, I saved both a full variant list and a list containing only variants of high or moderate predicted impact. All annotated variant lists from VEP were downloaded and filtered for unique variant changes by simultaneously removing rows with duplicates in all of the following VEP fields: Allele, Symbol, Gene, Amino Acid, and Codon. The resultant VEP-generated variant lists for the double-mutants only and stock-inclusive bedtools analyses are summarized in **Table 2-2**. The stock-inclusive analysis reduced the total variant list by 17.1%, but only reduced the moderate to high set of variants by 6.5%, indicating that much of the variants subtracted out in this analysis were low impact or modifier variants.

In terms of affected genes—some of which included multiple variants—narrowing this list by restricting it to the boundaries set by VRM led to a 26.9% reduction in the stock-inclusive analysis, compared to a 16.5% reduction when the stocks were not included. However, both lists of genes in the ROI with moderate to high predicted impact contained hundreds of hits for initial candidate consideration. With this in mind, I filtered again and narrowed the lists to include only high predicted impact hits. This narrowed the gene list to 35 in the double-mutant only set, and 18 in the stock-inclusive set (**Table 2-2**). Considering that this analysis of the round one strains included either only 9 recombinant genomes,

or 9 recombinant genomes and 2 stocks, we have demonstrated a powerful and efficient tool for obtaining a filtered list of genes with causative variant candidates for a spontaneous synthetic lethality phenotype.

Table 2-2. WGS third chromosome variant data

Data Set Analyzed	Total Variants	Total Variants (HIGH or MODERATE Predicted Impact)	Total HIGH to MOD. Genes	HIGH to MOD. Genes in ROI	HIGH only
Variants in lethals, not in viable	19,954	1166	558	466	35
Variants in lethals and *blm* stock, not in viable or *mus301* stock	16,548	1090	539	394 (139*)	18 (5*)
List of genes in ROI with HIGH predicted impact variants	*alpha-Est5, Ama, CG7448, CG11248, CR45951*				

*Values in parentheses represent genes remaining after applying new ROI, which was defined by VRM analysis of double mutants *bm-2-119* and *bm-2-161*.

The level of filtration achieved at the bioinformatics level was effective, but the ROI size of 20 Mbp remained large. In order to provide a more powerful filter with the ROI, I initiated a repeat of the crosses which were originally done to create the blm^{N1} $mus301^{288A}$ double mutants (discussed previously as round two of crosses). The availability of additional recombination events introduces the possibility for further refinement of the ROI using the same analysis shown in **Figure 2-5**. Because the ROI is centered around the third chromosome centromere, we do expect recombination events to be repressed in the this heterochromatic region (Denell & Keppy, 1979). I accounted for this by setting up

an increased number of single male crosses—and therefore a higher number of potential recombination events—in our second round of crosses. This second led to the creation of 12 new double mutants, as described previously. All of these were homozygous lethal.

However, with the generation of the ROI, it was not necessary to sequence all 12 new strains. I referred back to our VRM analysis and scanned the region around the left arm ROI boundary, which was defined by the *bm-6* strain (**Figure 2-5a, b**). I discovered a 39 bp indel located 116.7 kb downstream of the left arm 70B2 crossover in *bm-6*. This indel is present in the blm^{N1} stock and in the *bm-6* double mutant, but not in the $mus301^{288A}$ stock. I designed primers that flanked this indel, and did a PCR with these primers for each of the 12 new double mutant strains, using the *bm-6* as a positive (blm^{N1}-like) control. All double mutants genotyped as $mus301^{288A}$-like at this position (indel *not* present) were selected as candidates for sequencing, as these were the only recombinants demonstrating the potential to refine the ROI. Of the 12 new double mutant strains, only two fell into this category: *bm-2-119* and *bm-2-161*. I did not genotype the right arm ROI boundary, as it was specific to the *bm-14* viable strain, and we had no new viable strains to analyze. The WGS data from these new strains did, in fact, provide significant refinement to the ROI, thereby provided us with a smaller and more manageable list of candidate variants **(Table 2-2**, **Table A-2-1)**.

DISCUSSION

Helicases are key to maintaining genomic stability in the face of both endogenous and exogenous sources of DNA damage. Because damage can occur at any point of the cell cycle, organisms need to have a variety of repair pathways at their disposal that can respond to DNA damage during both gametogenesis and replication. Several DNA helicases act in more than one of these pathways, and display functions in both meiotic and mitotic repair. Evidence in multiple model system supports BLM and HELQ as two helicases that fit this description. In *D. melanogaster*, Blm has been well characterized as a key mediator of both recombinational and non-recombinational repair (McVey et al., 2007). In its ability to promote SDSA via migration and disruption of D-loops, it can prevent mitotic crossovers and repair mitotic DSB via HDR. It can also achieve repair after second-end capture by the dissolution of dHJ junction structures in cooperation with Top3α. BLM provides an additional safeguard to replication by unwinding G4 DNA and promoting replication fork regression and fork restart. Recently, a helicase-independent role was established for DmBlm in the gene conversion tract length in HDR, which may be related to results showing that human BLM interacted with DNA2 in a helicase-independent mechanism to promote end resection (Ertl et al., 2017; Nimonkar et al., 2011). BLM's presence throughout the cell cycle, and at points both early and late in repair pathways, has solidified its status as a genomic caretaker. While DmHelQ is not as well studied, evidence of the roles of HELQ in repair and genomic maintenance in multiple other systems has underlined its importance, despite the fact that it seems to be

more specialized than BLM. Both genes have been shown to be important to early development in *D. melanogaster*, and both can be indicators of cancer risk in humans. This makes BLM and HELQ important research targets, both in model systems and in clinical labs. In this study, I used two well-characterized null alleles of DmBlm (blm^{N1}) and DmHelQ ($mus301^{288A}$) to generate double mutants and seek novel interactions and implications for the roles of these helicases in repair and genome integrity (**Figure 2-1**).

I set up 320 recombinant single male *D. melanogaster* crosses (**Figure 2-6**) to generate double mutant strains. This was accomplished via two separate rounds of crosses: round one generated 9 total double mutants, while round two generated a total of 12 double mutants. I discovered and analyzed a synthetic lethality in these $blm^{N1}mus301^{288A}$ double mutants. Interestingly, the synthetic lethality was not present in all of them. Out of the 21 total double mutant strains obtained, 19 were synthetically homozygous lethal and 2 were homozygous viable, with lethality occurring between the third instar and pharate adult stages of development. Lethality at this stage of development may argue for an interaction based on repair during replication, as differentiation in multiple tissues at this stage poses significant demands on the replicative machinery and DNA damage checkpoint proteins. This has been demonstrated through studies in *D. melanogaster* which have underlined the importance of repair and checkpoint genes during larval and pupal stages (Brodsky, Sekelsky, Tsang, Hawley, & Rubin, 2000; Gorski et al., 2004). However, this alone does not allow for variants in other genes, such as those involved in *D. melanogaster* growth and

development, to be removed from consideration as causative in a larval lethality phenotype. To account for any initial bias toward certain types of genes during the screening of candidate variants for validation, all full data sets and variant lists are kept in my WGS workflow. These full data sets, which also include modifier variants and variants of low predicted impact, can always be mined for other genes of interest should validation of initial candidates fail. Because synthetic lethality was not present in all double mutants, we proposed that a single causative mutation had crossed over with either *blm* or *mus301* to cause the larval lethality in the 9 original affected strains (**Figure 2-3**). This was also based on our observation that the viable strain, *bm-14*, was able to produce homozygous adults at Mendelian ratios (**Table 2-1**).

Comment [d9]: This is my attempt to acknowledge the possibility for us to find other types of genes.

To find a causative mutation, I conducted WGS and applied a bioinformatics pipeline to the 9 double mutant strains obtained in the first round of crosses. Using this WGS data, I applied a VRM approach (**Figure 2-5**) to determine a ROI for a causative mutation. I successfully mapped crossover events in all 9 double mutant strains, and determined a 20 Mbp ROI (from cytogenic map location 70B2 through 85B7). Because the third chromosome in *D. melanogaster* is so large (over 50 Mbp in total) crossover events on the left arm and right arm can be considered independently. Out of total of 18 arms analyzed, I observed 3 double crossovers, 13 single crossovers, and 1 arm without crossovers (**Figure 2-5**). This data supported the suppression of crossover events near the centromere (Denell & Keppy, 1979). However, because the ROI is relatively centered around the centromere and remained large after sequencing of the 9

double mutants from round one, I submitted round two double mutants for sequencing, with the intention of attempting to refine the ROI. For the double mutants from round two, instead of sequencing each one, I used information from our initial ROI to find a blm^{N1}-like 39 bp indel downstream of the left ROI boundary. Using indel flanking primers, I determined that only 2 of the 12 round two double mutants, (*bm-2-119* and *bm-2-161*), were $mus301^{288A}$-like at this location. VRM analysis of *bm-2-119* moved the left arm ROI boundary inward by a significant amount, from 70B2 to 78C9 (**Figure 2-5b**). VRM analysis of the *bm-2-161* lethal revealed a left arm *mus301*-like to *blm*-like crossover that was downstream of 70B2, but upstream of 78C9, and therefore did not affect the final left arm ROI boundary.

The WGS data was also used to generate a variant list (**Table 2-2**), to which the previously-described ROI was applied. I achieved this using the bioinformatics tool called bedtools. I used this tool to overlap the variant calls in the lethal isolates, from which I subtracted variant calls in the viable isolate. Because VRM analysis revealed that a causative mutant likely originated in the blm^{N1} mutant chromosome, I did a second analysis where the lethal set was also overlapped with blm^{N1}variant calls, and the subsequent subtraction included variant calls from the *bm-14* viable double mutant and the $mus301^{288A}$ mutant. This second analysis gave the list representing the total remaining variants that were present in the blm^{N1} stock and *all* lethal double mutants, but not in the *bm-14* viable double mutant *or* the $mus301^{288A}$ mutant stock (**Table 2-2**). Fortunately, the information gained from WGS of the round two *bm-2-119* and *bm-2-161* double

mutants tightened the ROI for the causative variant, and thereby narrowed the original list of candidates for validation (**Table 2-2, Table A-2-1**). However, this narrowed list also resulted in the removal of one particular candidate variant, which was my top candidate for validation from the original list, from further consideration (discussed in Appendix). Therefore, further analysis of the refined list, followed by validation of the candidate variants within the genes in this list, will be necessary before the nature of the genetic interactions and corresponding larval lethality phenotype in the blm^{N1} $mus301^{288A}$ double mutants discussed here can be understood. However, the level of bioinformatic filtration achieved in this complex genetic background was drastic, and demonstrated the utility of our WGS and bioinformatics pipeline in this project.

Comment [SBT10]: Added in about new data and new list…I refer to Appendix.

The information about the causative variant gained from ROI was also valuable in providing support for a potentially novel interaction. Larval stage synthetic lethality has been identified in blm^{N1} mutants in the past. This interaction was shown with several structure-specific endonucleases which are important to the resolution of recombination intermediates in the absence of BLM (Andersen et al., 2011). However, because we believe our causative mutation originates on the blm^{N1} chromosome (**Figure 2-5**), the likelihood of a similar interaction being responsible for our observations is unlikely. Instead, we suspect that the interaction could relate to an interaction between DmBlm, DmHelQ, and our causative mutation, although we cannot rule out the possibility that the interaction may also be independent of DmBlm, and therefore only indicative of a relationship between DmHelQ and the causative mutation. However, given the

lack of knowledge currently available on DmHelQ, both of these possibilities are equally intriguing.

MATERIALS AND METHODS

Generation of double mutants in *Drosophila melanogaster*

D. melanogaster stocks and crosses were maintained in bottles or vials at 25°C on standard cornmeal agar media. The mutant alleles for *blm* and *mus301* were both generated via imprecise P-element excision (Melissa D. Adams & Sekelsky, 2002). The P-element used to create the blm^{N1} allele was $P\{EPgy2\}mus309^{EY0374}$ (McVey et al., 2007). The P-element used to create the $mus301^{288A}$ allele was $P\{SUPor\text{-}P\}mus301^{KG09098}$. Single mutant stocks used for initial crosses were $mus301^{288A}$/*TM6B, Tb, Hu* and blm^{N1}/*TM6B, Tb, Hu*. The crossing scheme used to obtain third chromosome blm^{N1} $mus301^{288A}$ mutants is shown (**Figure 2-6**).

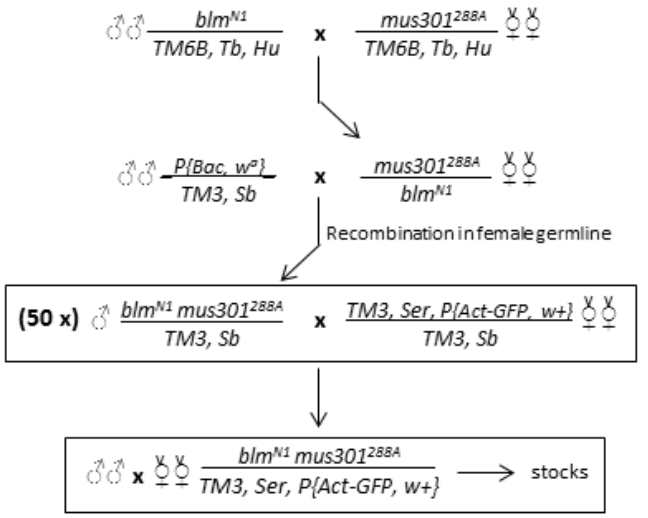


Figure 2-6. Cross scheme for generation of *blm^{N1} mus301^{288A}* double mutants. The (**50 x**) box refers to the 50 single male crosses set up at this stage. The final box represents sibling crosses between collected *blm^{N1} mus301^{288A}/TM3, Ser, P{Act-GFP}* males and females to establish double mutant strain stocks.

In the first round of crosses, a total of 50 single male crosses were set up. In the second round, 270 single male crosses were set up. Balancers used in the early steps of the second round cross scheme differed from those in **Figure 2-6**, but the final double mutant strain stocks were generated over the same *TM3, Ser, P{Act-GFP, w$^+$}* balancer used in the first round. For both the first and second round of single male crosses, each individual male was collected after 6 days for genotyping. Genomic DNA was extracted from these males using 50 µL Squishing Buffer (10 mM Tris-Cl, 25 mM NaCl, 1 mM EDTA) and Proteinase K (0.2 mg/mL). These were incubated at 37°C for 30 minutes, and 95°C for 2 min to inactivate Proteinase K. Mutant alleles were confirmed in the males via PCR and gel electrophoresis. Deletion-flanking primers (all listed here in 5' to 3'

orientation) that were used to genotype the blm^{N1} allele: -164(F): TGAAGGGTGGACCGACGGTC; 4811(R): GCCAGAATATCCAAGCGGAC. Deletion-flanking primers used to genotype $mus301^{288A}$: p^{OUT}(F): CCGCGGCCGCGGACCACCTTATGTTATTTC; 3586(R): CATGTTCCAGGTACCACACG.

Characterization of double mutant lethality phenotype and timing

Fly heterozygote/homozygote counts for double mutant strains were made over a period of 7 days post enclosure. Counts (**Table 2-1**) were primarily made from bottles, but counts for *bm-6* and *bm-16* were made from multiple vials and summed. Values for percent homozygotes eclosing was calculated as follows: (total homozygotes/total flies eclosed)*100%. The homozygous viable strain produced homozygotes at Mendelian ratios (24.51%) (**Table 2-1**). Round two double mutant isolate counts for homozygous lethality/viability were made similarly, but in vials over a period of 10 days. All double mutant strains were generated and maintained over either a *TM3, Ser, P{Act-GFP, w^+}* balancer, or, in the case of strain *bm-16*, a *TM6B, P{Act-GFP, w^+}* balancer (**Figure 2-6**). With these balancers, larvae could be screened for presence or absence of GFP to identify heterozygotes (GFP^+) and homozygotes (GFP^-). For each round one double mutant strain, heterozygous siblings were crossed and set up in grape plate agar cages. Grape agar plates in these cages were changed out every 24 hours for 3 days, for a total of 3 replicate grape agar plates for each strain. Larval and pupal

development was observed on grape agar plates over a period of 14 to 21 days, and observations were compiled to determine timing of lethality (**Figure 2-3**). For round two double mutant strains, lethality timing was confirmed as consistent with round one double mutant strains by making observations of grape agar plates and by observing GFP- pupae in vial under a fluorescent microscope.

Genomic DNA Extraction from whole flies and larvae

For each blm^{N1} $mus301^{288A}$ double mutant strain, heterozygous siblings were crossed and allowed to lay eggs on grape-juice agar plates for 24 hours. For each strain, 50 homozygous (GFP-) late second instar and third instar larvae were collected from grape-juice agar plates and rinsed with deionized water, then frozen down at -80°C. Larvae were homogenized in 600 µL DNA extraction lysis buffer (10mM Tris-HCl pH 8.0, 10 mM EDTA, 150 mM NaCl; plus addition of 30 µL of 10 mg/mL Proteinase K, 60 µL of 10% SDS, and 20 µL of 10 mg/mL RNaseA) and incubated at 65°C for 1 hour and allowed to cool to room temperature. Phenol:chloroform:isoamyl alcohol (25:24:1) was added to the lysis mixture (1:1), left on a rocking platform for 15 minutes, and spun down for 15 minutes at 14,000 at room temperature. This was repeated for one additional extraction with phenol:chloroform:isoamyl alcohol, after which a $1/5^{th}$ volume of 8M potassium acetate was added. This was followed by one final extraction with chloroform. DNA was precipitated from the supernatant with isopropanol, and incubated at -80°C for 30 minutes. The pellet was rinsed with 70% ethanol and

allowed to dry for 5-10 minutes. Pellets were re-suspended in 50 µL pico pure water. Initial quality control (QC) of genomic DNA was assessed by Nanodrop spectrophotometry and gel electrophoresis. For increased accuracy, DNA, RNA, and Protein quantification was assed using Qubit spectrophotometry. Residual RNA was removed by treatment with 10 mg/mL RNaseA followed by incubation at 37°C for at least 30 minutes, followed by clean-up using a Macherey-Nagel NucleoSpin Gel and PCR Clean-up kit.

Library Preparation and Illumina MiSeq WGS

Genomic DNA passing QC was diluted to 2.5 ng/µL in 20 µL. DNA libraries for MiSeq WGS were prepared according to the Illumina Nextera protocol for 2x250 paired-end sequencing runs (Illumina Inc., 2016). The Advanced Analytical (AATI) Fragment Analyzer (DNF-474 High Sensitivity NGS Fragment Analysis Kit) was used to assess fragment size distribution of DNA libraries before sequencing (Advanced Analytical Technologies, 2017). Paired-end (2x300) sequencing was done on an Illumina MiSeq. Most runs were confined to 3-4 indexed genomes per flowcell lane in order to balance coverage and efficiency. All index (index adapter) sequences used were from the Nextera

kit, and sequence information can be found on the “Index Sequences” page of the Nextera Library Prep Reference Guide (Illumina Inc., 2016).

WGS Data Analysis and filtration

Analysis via the Tufts University High Performance Compute Cluster (HPC): All WGS raw data was provided in the zipped (.gz) file format and transferred to the Tufts HPC using a free, open source file transfer program, FileZilla (https://filezilla-project.org/). We connected to the HPC remotely using another free, open source program: a SSH and Telnet client called PuTTY (http://www.putty.org/). The data manipulations described here were all accomplished via command line (Unix/Linux) and software modules available on the HPC network. WGS raw read .gz files were loaded onto our lab’s personal HPC directory, and decompressed to the FASTQ file format. If applicable, FASTQ files for genomes with multiple runs’ worth of data (either from multiple full WGS runs, or from a WGS run and a QC run) were concatenated. For each genome, FASTQ files for paired end reads (format: genome1_R1.fastq, genome1_R2.fastq) were processed via a paired-end alignment script (pe_align.sh) designed to generate variant call format (VCF) data. The script employed the following software for data manipulation, in order: Bowtie2 (alignments), SAMtools (quality filtration; SAM to binary format, BAM; sorting), Picard (remove PCR duplicates), SAMtools (index BAM files), GATK (create VCF from BAM). Finally, the script also selectively filtered final VCF output to

the third chromosome, and generated VCF files for the left arm, right arm, and entirety of that chromosome (format: genome1_3L.vcf, genome1_3R.vcf, genome1.vcf). The dm6 *Drosophila melanogaster* genome build was used as the reference genome for variant calling in the script. For generation of variant lists, the full chromosome VCF files were compared to each other using bedtools (version 2.26.0), which was included as a software module within the HPC. To overlap genomes (find similarities), we used bedtools-intersect. To subtract genomes (remove background), we used bedtools-subtract.

Post HPC Analyses and Filtration: After bedtools analysis, VCF files were downloaded via FileZilla, and were then annotated with the Ensembl Variant Effect Predictor (VEP) (http://www.ensembl.org/Tools/VEP) for *D. melanogaster* (BDGP6 genome assembly). Settings used were as follows (any settings not mentioned were not selected). Identifiers: Gene symbol, CCDS, Protein, Uniprot, HGVS, $CSN^{(p)}$. Frequency data (find co-located known variants): Yes. Miscellaneous: Transcript biotype, Protein domains, Exon and intron numbers, Transcript support level, APPRIS, Identify canonical transcripts, Upstream/Downstream distance (bp) – 5000 bp. After obtaining results from VEP, annotated variant information was downloaded as a TXT file. Within VEP, we also filtered results down to a list of variants with high and moderate predicted impact, and downloaded this information as a separate TXT file. To visualize variant lists, TXT files were viewed, manipulated, and saved as MS Excel workbooks. Any exact duplicate rows were removed from these files, Subsequently Allele, Gene, Symbol, Amino Acid, and Codon were removed in

order to create a list of unique variant hits. From this data, we generated gene lists from which to seek out interesting candidates for causative variants.

VCF Recombination mapping and region-of-interest (ROI)-based genotyping

Single-genome third chromosome VCF files (genome1.vcf) generated from our script (pe_align.sh) were used for recombination mapping. In order to visualize the location of the crossover events that had occurred in each blm^{N1} $mus301^{288A}$ double mutant strain, we used the Integrative Genomics Viewer (http://software.broadinstitute.org/software/igv/). Single genome VCF files for each double mutant strain were loaded into the IGV software individually. Each double mutant VCF file was viewed between the VCF files for the original blm^{N1} and $mus301^{288A}$ stock genomes. Crossover events were identified as the locations where the pattern of variant calls in the double mutant switched from resembling the pattern in blm^{N1} to resembling that of $mus301^{288A}$, and vice versa. Resulting recombination maps from each individual genome were analyzed together in order to find a region where all lethal strain genomes were alike, yet also differed from the viable strain genome (**Figure 2-5**). This region became our region of interest (ROI) for potentially causative variants, and was used to narrow the post bedtools variant lists discussed previously.

VRM data was also used to determine which of the 12 round two double mutant strains should be sequenced to refine the ROI. We genotyped each of these

strains for a 39 bp indel near the left ROI boundary that was present in the *blm*^{N1} sequence. We used PCR with indel-flanking primers to genotype the 12 round two double mutant strains as *blm*^{N1}-like or *mus301*^{288A}-like and selected only strains that genotyped as *mus301*^{288A}-like (and therefore potentially ROI refining) for subsequent sequencing.

ACKNOWLEDGEMENTS

I would like to thank Mitch McVey for experimental guidance, scientific insight, and for helpful reading and comments for this manuscript. A special thanks is due to Erik Dopman, whose suggestions related to recombination mapping and ROI genotyping were extremely valuable to this work. Sergei Mirkin and Stephen Fuchs also provided valuable guidance for this project. This research was supported by grant P01GM105473 from the NIH.

REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K A., Howland, T J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., Woodage T., Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X.,

Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., Venter, J. C. (2000). The genome sequence of Drosophila melanogaster. *Science*, *287*(5461), 2185–2195.

Adams, M. D., McVey, M., & Sekelsky, J. J. (2003). Drosophila BLM in double-strand break repair by synthesis-dependent strand annealing. *Science*, *299*(5604), 265–267.

Adams, M. D., & Sekelsky, J. J. (2002). From sequence to phenotype: reverse genetics in Drosophila melanogaster. *Nature Reviews Genetics*, *3*(3), 189–198.

Adelman, C. A., & Boulton, S. J. (2010, September 10). Metabolism of postsynaptic recombination intermediates. *FEBS Letters*, *584*(17), 3709-3716.

Adelman, C. A., Lolo, R. L., Birkbak, N. J., Murina, O., Matsuzaki, K., Horejsi, Z., Parmar, K., Borel, V., Skehel, J. M., Stamp, G., D'Andrea, A., Sartori, A. A., Swanton, C., Boulton, S. J. (2013). HELQ promotes RAD51 paralogue-dependent repair to avert germ cell loss and tumorigenesis. *Nature*, *502*(7471), 381–384.

Advanced Analytical Technologies. (2017). Fragment Analyzer Automated CE System: Quick Start Guide - 12 Capillary.

Andersen, S. L., Kuo, H. K., Savukoski, D., Brodsky, M. H., & Sekelsky, J. (2011). Three structure-selective endonucleases are essential in the absence of BLM helicase in Drosophila. *PLoS Genetics*, *7*(10), e1002315.

Beagan, K., Armstrong, R. L., Witsell, A., Roy, U., Renedo, N., Baker, A. E., Scharer, O., McVey, M. (2017). Drosophila DNA polymerase theta utilizes both helicase-like and polymerase domains during microhomology-mediated end joining and interstrand crosslink repair. *PLoS Genetics*, *13*(5), e1006813.

Boyd, J. B., Golino, M D., Shaw, K. E. S., Osgood, C. J., Green, M. M. (1981). Third-chromosome mutagen-sensitive mutants of Drosophila melanogaster. *Genetics*, *97*(3–4), 607–623.

Brodsky, M. H., Sekelsky, J. J., Tsang, G., Hawley, R. S., & Rubin, G. M. (2000). mus304 encodes a novel DNA damage checkpoint protein required during Drosophila development. *Genes and Development*, *14*(6), 666–678.

Brosh, R. M. (2013). DNA helicases involved in DNA repair and their roles in cancer. *Nature Reviews Cancer*, *13*(8), 542–558.

Chan, S. H., Yu, A. M., & McVey, M. (2010). Dual Roles for DNA Polymerase Theta in Alternative End-Joining Repair of Double-Strand Breaks in Drosophila. *PLoS Genetics*, *6*(7), e1001005.

Cheok, C. F., Wu, L., Garcia, P. L., Janscak, P., & Hickson, I. D. (2005). The Bloom's syndrome helicase promotes the annealing of complementary single-stranded DNA. *Nucleic Acids Research*, *33*(12), 3932–3941.

Davalos, A. R., & Campisi, J. (2003). Bloom syndrome cells undergo p53-dependent apoptosis and delayed assembly of BRCA1 and NBS1 repair complexes at stalled replication forks. *Journal of Cell Biology*, *162*(7), 1197–1209.

Davalos, A. R., Kaminker, P., Hansen, R. K., & Campisi, J. (2004). ATR and ATM-dependent movement of BLM helicase during replication stress ensures optimal ATM activation and 53BP1 focus formation. *Cell Cycle*, *3*(12), 1579–1586.

de Renty, C., & Ellis, N. A. (2017, January 1). Bloom's syndrome: Why not premature aging?: A comparison of the BLM and WRN helicases. *Ageing Research Reviews*, *33*, 36–51.

Denell, R. E., & Keppy, D. O. (1979). The nature of genetic recombination near the third chromosome centromere of Drosophila melanogaster. *Genetics*, *93*(1), 117–130.

Dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., Emmert, D. B., Gelbart, W. M., Brown, N. H., Kaufman, T., Werner-Washburne, M., Cripps, R., Broll, K., Gramates, L. S., Falls, K., Matthews, B. B., Russo, S., Zhou, P., Zytkovicz, M., Adryan, B., Attrill, H., Costa, M., Marygold, S., McQuilton, P., Millburn, G., Ponting, L., Stefancsik, R., Tweedie, S., Grumbling, G. (2015). FlyBase: Introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, *43*(D1), D690–D697.

Ellis, N. A., Groden, J., Ye, T. Z., Straughen, J., Lennon, D. J., Ciocci, S., Proytcheva, M., German, J. (1995). The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell*, *83*(4), 655–666.

Ellis, N. A., Proytcheva, M., Sanz, M. M., Ye, T. Z., & German, J. (1999). Transfection of BLM into cultured bloom syndrome cells reduces the sister-chromatid exchange rate toward normal. *American Journal of Human Genetics*, *65*(5), 1368–1374.

Ertl, H. A., Russo, D. P., Srivastava, N., Brooks, J. T., Dao, T. N., & LaRocque, J. R. (2017). The Role of Blm Helicase in Homologous Recombination, Gene Conversion Tract Length, and Recombination Between Diverged Sequences in Drosophila. *Genetics*, *207*(3), 923–933.

Ghabrial, A., & Schüpbach, T. (1999). Activation of a meiotic checkpoint regulates translation of Gurken during Drosophila oogenesis. *Nature Cell Biology*, *1*(6), 354–357.

González-Reyes, A., Elliott, H., & St Johnston, D. (1997). Oocyte determination and the origin of polarity in Drosophila: the role of the spindle genes. *Development*, *124*(24), 4927–4937.

Gorski, M. M., Romeijn, R. J., Eeken, J. C. J., De Jong, A. W. M., Van Veen, B. L., Szuhai, K., Mullenders, L. H., Ferro, W., Pastink, A. (2004). Disruption of Drosophila Rad50 causes pupal lethality, the accumulation of DNA double-strand breaks and the induction of apoptosis in third instar larvae. *DNA Repair*, *3*(6), 603–615.

Grabarz, A., Guirouilh-Barbat, J., Barascu, A., Pennarun, G., Genet, D., Rass, E., Germann, S. M., Bertrand, P., Hickson, I. D., Lopez, B. S. (2013). A Role for BLM in Double-Strand Break Repair Pathway Choice: Prevention of CtIP/Mre11-Mediated Alternative Nonhomologous End-Joining. *Cell Reports*, *5*(1), 21–28.

Gramates, L. S., Marygold, S. J., Dos Santos, G., Urbano, J. M., Antonazzo, G., Matthews, B. B., Rey, A. J., Tabone, C. J., Crosby, M. A., Emmert, D. B., Falls, K., Goodman, J. L., Hu, Y., Ponting, L., Schroeder, A. J., Strelets, V. B., Thurmond, J., Zhou, P., Perrimon, N., Gelbart, S. R., Extavour, C., Broll, K., Zytkovicz, M., Brown, N. H., Attrill, H., Costa, M., Fexova, S., Jones, T., Larkin, A., Millburn, G. H., Staudt, N., Kaufman, T., Grumbling, G. B., Cripps, R., Werner-Washburne, M., Baker, P. (2017). FlyBase at 25: Looking to the future. *Nucleic Acids Research*, *45*(D1), D663–D671.

Gruber, S. B. (2002). BLM Heterozygosity and the Risk of Colorectal Cancer. *Science*, *297*(5589), 2013–2013.

Hamdi, Y., Soucy, P., Adoue, V., Michailidou, K., Canisius, S., Lemaçon, A., Droit, A., Andrulis, I. L., Anton-Culver, H., Arndt, V., Baynes, C., Blomqvist, C., Bogdanova, N. V., Bojesen, S. E., Bolla, M. K., Bonanni, B., Borresen-Dale, A., Brand, J. S., Brauch, H., Brenner, H., Broeks, A., Burwinkel, B., Chang-Claude, J., Collaborators, NBCS., Couch, F. J., Cox, A., Cross, S. S., Czene, K., Darabi, H., Dennis, J., Devilee, P., Dörk, T., Dos-Santos-Silva, I., Eriksson, M., Fasching, P. A., Figueroa, J., Flyger, H., García-Closas, M., Giles, G. G., Goldberg, M. S., González-Neira, A., Grenaker-Alnæs, G., Guénel, P., Haeberle, L., Haiman, C. A., Hamann, U.,

Hallberg, E., Hooning, M. J., Hopper, J. L., Jakubowska, A., Jones, M., Kabisch, M., Kataja, V., Lambrechts, D., Marchand, L. L., Lindblom, A., Lubinski, J., Mannermaa, A., Maranian, M., Margolin, S., Marme, F., Milne, R. L., Neuhausen, S. L., Nevanlinna, H., Neven, P., Olswold, C., Peto, J., Plaseska-Karanfilska, D., Pylkäs, K., Radice, P., Rudolph, A., Sawyer, E. J., Schmidt, M. K., Shu, X., Southey, M. C., Swerdlow, A., Tollenaar, R. A. E. M., Tomlinson, I., Torres, D., Truong, N., Vachon, C., Van Den Ouweland, A. M.W., Wang, Q., Winqvist, R., Zheng, W., Benitez, J., Chenevix-Trench, G., Easton, A., Pastinen, T., Nord, S., Simard, J. (2016). Association of breast cancer risk with genetic variants showing differential allelic expression: Identification of a novel breast cancer susceptibility locus at 4q21. *Oncotarget*, *7*(49), 80140–80163.

Han, X., Zhao, L., & Li, X. (2016). HELQ in cancer and reproduction. *Neoplasma*, *63*(6), 825–835.

Holloway, J. K., Morelli, M. A., Borst, P. L., & Cohen, P. E. (2010). Mammalian BLM helicase is critical for integrating multiple pathways of meiotic recombination. *Journal of Cell Biology*, *188*(6), 779–789.

Illumina Inc. (2016). Nextera Library Prep Reference Guide. Sample Preparation Guide, (January), 1–28. https://doi.org/RS-122-2001; RS-122-2002

Ira, G., Malkova, A., Liberi, G., Foiani, M., & Haber, J. E. (2003). Srs2 and Sgs1 – Top3 Suppress Crossovers during Double-Strand Break Repair in Yeast. *Cell*, *115*, 401–411.

Johnson-Schlitz, D., & Engels, W. R. (2006). Template disruptions and failure of double Holliday junction dissolution during double-strand break repair in Drosophila BLM mutants. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(45), 16840–16845.

Karow, J. K., Constantinou, A., Li, J. L., West, S. C., & Hickson, I. D. (2000). The Bloom's syndrome gene product promotes branch migration of holliday junctions. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(12), 6504–6508.

Kohn, K. W., Spears, C. L., & Doty, P. (1966). Inter-strand crosslinking of DNA by nitrogen mustard. *Journal of Molecular Biology*, *19*(2), 266–288.

Kusano, K. (2001). Sterility of Drosophila with Mutations in the Bloom Syndrome Gene--Complementation by Ku70. *Science*, *291*(5513), 2600–2602.

MacHwe, A., Karale, R., Xu, X., Liu, Y., & Orren, D. K. (2011). The Werner and Bloom syndrome proteins help resolve replication blockage by converting

(regressed) Holliday junctions to functional replication forks. *Biochemistry*, *50*(32), 6774–6788.

Maréchal, A., & Zou, L. (2013). DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harbor Perspectives in Biology*, *5*(9), a012716.

Marini, F., & Wood, R. D. (2002). A human DNA helicase homologous to the DNA cross-link sensitivity protein Mus308. *Journal of Biological Chemistry*, *277*(10), 8716–8723.

McVey, M., Andersen, S. L., Broze, Y., & Sekelsky, J. (2007). Multiple functions of drosophila BLM helicase in maintenance of genome stability. *Genetics*, *176*(4), 1979–1992.

McVey, M., Larocque, J. R., Adams, M. D., & Sekelsky, J. J. (2004). Formation of deletions during double-strand break repair in Drosophila DmBlm mutants occurs after strand invasion. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(44), 15694–15699.

Mendoza, O., Bourdoncle, A., Boulé, J. B., Brosh, R. M., & Mergny, J. L. (2016, March 18). G-quadruplexes and helicases. *Nucleic Acids Research*, *44*(5), 1989–2006.

Mohaghegh, P., Karow, J. K., Brosh, R. M., Bohr, V. A., & Hickson, I. D. (2001). The Bloom's and Werner's syndrome proteins are DNA structure-specific helicases. *Nucleic Acids Research*, *29*(13), 2843–2849.

Muzzini, D. M., Plevani, P., Boulton, S. J., Cassata, G., & Marini, F. (2008). Caenorhabditis elegans POLQ-1 and HEL-308 function in two distinct DNA interstrand cross-link repair pathways. *DNA Repair*, *7*(6), 941–950.

Nakayama, H., Nakayama, K., Nakayama, R., Irino, N., Nakayama, Y., & Hanawalt, P. C. (1984). Isolation and genetic characterization of a thymineless death-resistant mutant of Escherichia coli K12: Identification of a new mutation (recQ1) that blocks the RecF recombination pathway. *Molecular & General Genetics*, *195*(3), 474–480.

Nimonkar, A. V., Genschel, J., Kinoshita, E., Polaczek, P., Campbell, J. L., Wyman, C., Modrich, P., Kowalczykowski, S. C. (2011). BLM-DNA2-RPA-MRN and EXO1-BLM-RPA-MRN constitute two DNA end resection machineries for human DNA break repair. *Genes and Development*, *25*(4), 350–362.

Nimonkar, A. V, Ozsoy, A. Z., Genschel, J., Modrich, P., & Kowalczykowski, S. C. (2008). Human exonuclease 1 and BLM helicase interact to resect DNA

and initiate DNA repair. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(44), 16906–16911.

Payne, M., & Hickson, I. D. (2009). Genomic instability and cancer: lessons from analysis of Bloom's syndrome. *Biochemical Society Transactions*, *37*(Pt 3), 553–559.

Pelttari, L. M., Kinnunen, L., Kiiski, J. I., Khan, S., Blomqvist, C., Aittomäki, K., & Nevanlinna, H. (2016). Screening of HELQ in breast and ovarian cancer families. *Familial Cancer*, *15*(1), 19–23.

Roote, J., & Prokop, A. (2013). How to Design a Genetic Mating Scheme: A Basic Training Package for Drosophila Genetics. *G3: Genes|Genomes|Genetics*, *3*(2), 353–358.

Rosales-Nieves, A. E., & González-Reyes, A. (2014, April 1). Genetics and mechanisms of ovarian cancer: Parallels between Drosophila and humans. *Seminars in Cell and Developmental Biology*, *28*, 104–109.

Sengupta, S., Linke, S. P., Pedeux, R., Yang, Q., Farnsworth, J., Garfield, S. H., Valerie, K., Shay, J. W., Ellis, N. A., Wasylyk, B., Harris, C. C. (2003). BLM helicase-dependent transport of p53 to sites of stalled DNA replication forks modulates homologous recombination. *EMBO Journal*, *22*(5), 1210–1222.

Sengupta, S., Robles, A. I., Linke, S. P., Sinogeeva, N. I., Zhang, R., Pedeux, R., Ward, I. M., Celeste, A., Nussenzweig, A., Chen, J., Halazonetis, T. D., Harris, C. C. (2004). Functional interaction between BLM helicase and 53BP1 in a Chk1-mediated pathway during S-phase arrest. *Journal of Cell Biology*, *166*(6), 801–813.

Sheu, Y.-J., Kinney, J. B., Lengronne, A., Pasero, P., & Stillman, B. (2014). Domain within the helicase subunit Mcm4 integrates multiple kinase signals to control DNA replication initiation and fork progression. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(18), E1899–E1908.

Siede, W. (2014). DNA Interstrand Crosslink Repair. *eLS*, 1–10.

Stolk, L., Perry, J. R. B., Chasman, D. I., He, C., Mangino, M., Sulem, P., Barbalic, M., Broer, L., Byrne, E. M., Ernst, F., Esko, T., Franceschini, N., Gudbjartsson, D. F., Hottenga, J., Kraft, P., McArdle, P. F., Porcu, E., Shin, S., Smith, A. V., van Wingerden, S., Zhai, G., Zhuang, W. V., Albrecht, E., Alizadeh, B. Z., Aspelund, T., Bandinelli, S., Lauc, L., Beckmann, J. S., Boban, M., Boerwinkle, E., Broekmans, F. J., Burri, A., Campbell, H., Chanock, S. J., Chen, C., Cornelis, M. C., Corre, T., Coviello, A. D.,

D'Adamo, P., Davies, G., de Faire, U., de Geus, E. J. C., Deary, I. J., Dedoussis, G. V. Z., Deloukas, P., Ebrahim, S., Eiriksdottir, G., Emilsson, V., Eriksson, J. G., Fauser, B. C. J. M., Ferreli, L., Ferrucci, L., Fischer, K., Folsom, A. R., Garcia, M. E., Gasparini, P., Gieger, C., Glazer, N., Grobbee, D. E., Hall, P., Haller, T., Hankinson, S. E., Hass, M., Hayward, C., Heath, A. C., Hofman, A., Ingelsson, E., Janssens, A. C. J. W., Johnson, A. D., Karasik, D., Kardia, S. L. R., Keyzer, J., Kiel, D. P., Kolcic, I., Kutalik, Z., Lahti, J., Lai, S., Laisk, T., Laven, J. S. E., Lawlor, D. A., Liu, J., Lopez, L. M., Louwers, Y. V., Magnusson, P. K. E., Marongiu, M., Martin, N. G., Klaric, I. M., Masciullo, C., McKnight, B., Medland, S. E., Melzer, D., Mooser, V., Navarro, P., Newman, A. B., Nyholt, D. R., Onland-Moret, N. C., Palotie, A., Paré, G., Parker, A. N., Pedersen, N. L., Peeters, P. H. M., Pistis, G., Plump, A. S., Polasek, O., Pop, V. J. M., Psaty, B. M., Räikkönen, K., Rehnberg, E., Rotter, J. I., Rudan, I., Sala, C., Salumets, A., Scuteri, A., Singleton, A., Smith, J. A., Snieder, H., Soranzo, N., Stacey, S. N., Starr, J. M., Stathopoulou, M. G., Stirrups, K., Stolk, R. P., Styrkarsdottir, U., Sun, Y. V., Tenesa, A., Thorand, B., Toniolo, D., Tryggvadottir, L., Tsui, K., Ulivi, S., van Dam, R. M., van der Schouw, Y. T., van Gils, C. H., van Nierop, P., Vink, J. M., Visscher, P. M., Voorhuis, M., Waeber, G., Wallaschofski, H., Wichmann, H. E., Widen, E., Wijnands-van Gent, C. J. M., Willemsen, G., Wilson, J. F., Wolffenbuttel, B. H. R., Wright, A. F., Yerges-Armstrong, L. M., Zemunik, T., Zgaga, L., Zillikens, M. C., Zygmunt, M., LifeLines Cohort Study, Arnold, A. M., Boomsma, D. I., Buring, J. E., Crisponi, L., Demerath, E. W., Gudnason, V., Harris, T. B., Hu, F. B., Hunter, D. J., Launer, L. J., Metspalu, A., Montgomery, G. W., Oostra, B. A., Ridker, P. M., Sanna, S., Schlessinger, D., Spector, T. D., Stefansson, K., Streeten, E. A., Thorsteinsdottir, U., Uda, M., Uitterlinden, A. G., van Duijn, C. M., Völzke, H., Murray, A., Murabito, J. M., Visser, J. A., Lunetta, K. L. (2012). Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nature Genetics*, *44*(3), 260–268.

Suzuki, T., Yasu, M., & Honma, M. (2016). Mutator Phenotype and DNA Double-Strand Break Repair in BLM Helicase-Deficient Human Cells. *Molecular and Cellular Biology*, *36*(23), 2877–2889.

Tafel, A. A., Wu, L., & McHugh, P. J. (2011). Human HEL308 localizes to damaged replication forks and unwinds lagging strand structures. *Journal of Biological Chemistry*, *286*(18), 15832–15840.

Takata, K., Reh, S., Tomida, J., Person, M. D., & Wood, R. D. (2013). Human DNA helicase HELQ participates in DNA interstrand crosslink tolerance with ATR and RAD51 paralogs. *Nature Communications*, *4*, 2338.

Trowbridge, K., McKim, K., Brill, S. J., & Sekelsky, J. (2007). Synthetic lethality of drosophila in the absence of the MUS81 endonuclease and the DmBlm helicase is associated with elevated apoptosis. *Genetics*, *176*(4), 1993–2001.

Ui, A., Satoh, Y., Onoda, F., Miyajima, A., Seki, M., & Enomoto, T. (2001). The N-terminal region of Sgs1, which interacts with Top3, is required for complementation of MMS sensitivity and suppression of hyper-recombination in sgs1 disruptants. *Molecular Genetics and Genomics*, *265*(5), 837–850.

Van Brabant, A. J., Ye, T., Sanz, M., German, J. L., Ellis, N. A., & Holloman, W. K. (2000). Binding and melting of D-loops by the Bloom syndrome helicase. *Biochemistry*, *39*(47), 14617–14625.

Ward, J. D., Barber, L. J., Petalcorin, M. I., Yanowitz, J., & Boulton, S. J. (2007). Replication blocking lesions present a unique substrate for homologous recombination. *The EMBO Journal*, *26*(14), 3384–3396.

Ward, J. D., Muzzini, D. M., Petalcorin, M. I. R., Martinez-Perez, E., Martin, J. S., Plevani, P., Cassata, G., Marini, F., Boulton, S. J. (2010). Overlapping Mechanisms Promote Postsynaptic RAD-51 Filament Disassembly during Meiotic Double-Strand Break Repair. *Molecular Cell*, *37*(2), 259–272.

Woodman, I. L., Brammer, K., & Bolt, E. L. (2011). Physical interaction between archaeal DNA repair helicase Hel308 and Replication Protein A (RPA). *DNA Repair*, *10*(3), 306–313.

Wu, L., Davies, S. L., Levitt, N. C., & Hickson, I. D. (2001). Potential Role for the BLM Helicase in Recombinational Repair via a Conserved Interaction with RAD51. *Journal of Biological Chemistry*, *276*(22), 19375–19381.

Wu, L., & Hickson, I. D. (2003). The Bloom's syndrome helicase suppresses crossing over during homologous recombination. *Nature*, *426*(6968), 870–874.

You, Z., & Masai, H. (2017). Potent DNA strand annealing activity associated with mouse Mcm2∼7 heterohexameric complex. *Nucleic Acids Research*, *45*(11), 6494–6506.

APPENDIX

Initial analysis from round one double mutant strains identifies interesting dicistronic locus

The original list of high impact variants from the initial analysis of WGS data is presented here. This analysis did not include the two double mutants (*bm-2-119* and *bm-2-161* from the second round of crosses to obtain double mutants). One locus of particular interest from this initial analysis was *CG42374/CG9666*. Differential splicing of this dicistronic locus produces two protein products. Predicted model organism orthologs of *CG9666* are conserved from *C. elegans* to humans, and *CG9666* is predicted to have nucleic acid binding and methyltransferase activity (Gramates et al., 2017). Alternatively, *CG42374* has predicted involvement in DNA repair, and has orthologs in model metazoans from zebrafish to humans (Gramates et al., 2017).

The protein product has similarity to the sensor of ssDNA (SOSS) complex subunit C (SOSS-C). The SOSS complex can bind to ssDNA at the site of DNA DSB, and is likely to have an important role in repair and maintenance of genomic stability (Nam & Cortez, 2009). In human fibroblasts, the B component

of the heterotrimeric SOSS complex, hSSB1, was recruited to the sites of IR-induced DSBs, colocalizing with γ-H2AX (Richard et al., 2008). Finally, two more recent studies demonstrated that the SOSS complex could cooperate with Exo1 to promote end resection of dsDNA, and that hSSB1 interacted with BLM, both before the induction of ionizing radiation, and again 2-3 hours post-treatment (Croft et al., 2017; Yang et al., 2012). Together, these studies implicated *CG42374* as a high priority locus. The variant called in *CG42374/CG966*—which bedtools analysis assigns as present in all lethal isolates and the blm^{N1} stock, but not in the *bm-14* viable strain or the $mus301^{288A}$ stock—is a unique splice donor variant that eliminates the conserved GT splice site. The location of this variant has the potential to alter the splicing of both *CG42374* and *CG9666*. The WGS data for *bm-2-119* and *bm-2-161* removed this variant from consideration. Its cytogenic map location at 76A3 was no longer a part of the region of interest once the left arm boundary for was pushed inward to 78C9 by VCF recombination mapping analysis of *bm-2-119*. However, while no longer in consideration as causative in the context of the work described here, this may still be an interesting locus, and variant, to investigate further in the context of DNA repair and genomic integrity.

Loci with variants of moderate predicted-impact within the refined ROI

validation efforts have yet to take place, many of the genes containing variants of moderate predicted impact may also have to be considered downstream, depending on the results of validation efforts with the high-predicted impact variants. In order to account for this possibility, I have included a full list of the genes containing variants with moderate predicted impact below. As with any candidate variant, before any of the variants in the genes in this list are considered, sequencing alignments must also be viewed to filter out low quality calls and false positives from consideration. An example of this was described in Chapter 3 for the variant in *Claspin* (false positive) and the variant in *lodestar* (confirmed variant candidate).

Table A-2-1. List of genes in refined ROI with high and moderate impact variants

Gene symbol	Gene symbol	Gene symbol	Gene symbol	Gene symbol
Ac78C	*CG7133*	*CG11248*	*CG42337*	*MAGE*
alpha-Est10	*CG7148*	*CG11737*	*CG42564*	*mRpL1*
alpha-Est4	*CG7173*	*CG11762*	*CG43060*	*Neu2*
alpha-Est5	*CG7407*	*CG14463*	*CG43061*	*Nlg1*
alpha-Est8	*CG7414*	*CG14562*	*CG43254*	*Nopp140*
Als2	*CG7443*	*CG14563*	*CG43618*	*nxf4*
Ama	*CG7448*	*CG14564*	*CG45263*	*Or83c*
bel	*CG7519*	*CG14569*	*CG46026*	*Or85a*
Ccp84Ad	*CG7878*	*CG14570*	*CheA84a*	*Or85c*
Ccp84Ag	*CG7900*	*CG14572*	*CR45951*	*Or85d*
CD98hc	*CG8145*	*CG14573*	*Cyp313b1*	*Or85e*
Cdk12	*CG8159*	*CG14598*	*djl*	*Osi1*
Cenp-C	*CG8202*	*CG14608*	*DNApol-eta*	*Osi17*
CG1041	*CG8223*	*CG15186*	*DNApol-iota*	*p*
CG1091	*CG9626*	*CG17816*	*Dpck*	*pb*
CG1104	*CG9630*	*CG18249*	*eg*	*pch2*
CG1227	*CG9684*	*CG31259*	*Eip78C*	*ppk5*
CG1288	*CG9773*	*CG31463*	*Est-Q*	*puc*
CG1979	*CG10032*	*CG31482*	*gfzf*	*pyd*
CG1988	*CG10055*	*CG31496*	*Glg1*	*RacGAP84C*
CG2616	*CG10092*	*CG31544*	*gpp*	*rn*
CG2678	*CG10286*	*CG31560*	*Hr78*	*RpA-70*
CG2698	*CG10445*	*CG32436*	*Ilk*	*S1P*

CG2747		*CG10512*		*CG33288*		*Ir84a*	*sas*
CG2767		*CG10566*		*CG33290*		*lab*	*Syt4*
CG2943		*CG10581*		*CG34023*		*lap*	*unc-45*
CG3014		*CG10585*		*CG34127*		*lds (lodestar)*	*wa-cup*
CG5656		*CG11035*		*CG34384*		*mAChR-B*	

Cells shown in orange represent genes containing variants of high predicted impact. Some genes containing high impact variants also contain moderate impact variants. The full name of *lodestar* is given (symbol: *lds*), as it is discussed in Chapter 3.

Chapter 3

WGS and bioinformatics identify causative variant candidates for MMS hypersensitivity in mutants obtained via EMS-mutagenesis of Rev1-ΔCTD *Drosophila melanogaster*

Contributions to this work:

EMS mutagenesis and generation of mutant stocks: Varandt Khodaverdian, Hannah Slutsky, Natalie Danziger

Characterization and confirmation of MMS-hypersensitive mutants: Tokio Sano, Sarah Shnayder, Jake Cosgrove

Complementation testing: Tokio Sano, Sarah Shnayder, Jake Cosgrove

ABSTRACT

DNA Damage Tolerance (DDT) is a set of mechanisms that allows cells to synthesize past DNA lesions during replication, and thereby prevents the more detrimental impacts of the lesion, such as fork stalling or collapse. DDT is proposed to branch off into two main sub-pathways: a pathway dependent on specialized error-prone polymerases, called trans-lesion synthesis (TLS), and another recombination-based pathway called template-switching (TS). While some of the main TLS polymerases and the TLS mechanisms of repair have been well-characterized, the TS pathway remains poorly understood in comparison. In *Drosophila melanogaster*, we have previously gathered evidence that the translesion polymerase and scaffolding protein Rev1 is essential to both TLS and TS, and that loss of its C-terminal domain (CTD) removes its ability to recruit TLS polymerases such as Polζ. Therefore, without access to TLS, *Rev1*-Δ*CTD* flies should depend on TS to achieve DDT. In order to gain a deeper understanding of REV1 in DDT pathway choice in a model metazoan, and to identify other proteins important to TS in an unbiased manner, we conducted an ethyl-methane sulfonate (EMS)-mutagenesis screen in *Rev1*-Δ*CTD* background in

D. melanogaster. From this screen, we obtained 23 mutant strains that showed hyper-sensitivity to the alkylating agent methyl-methane sulfonate (MMS). Here, we discuss the WGS analysis of 3 of these mutants: one set of 2 non-complementing mutants (148 and 157), as well as one individual mutant (1396). We demonstrate that the application of a single bioinformatics and data analysis pipeline to each of these WGS data sets is sufficient to identify candidates for causative variants in each. In the non-complementing set, we identified a single gene, *polybromo*, with unique high-impact variants in each of the non-complementing mutants. In the single mutant, we identified an interesting candidate, *lodestar*, which also had a unique variant, and shows conservation with human HLTF. Further validation of these genes as causative could provide valuable, novel insight into the mechanisms and proteins involved in the TS sub-pathway of DDT.

INTRODUCTION

Rev1 and DNA Damage Tolerance (DDT) in eukaryotes

During an organism's development, growth and differentiation depend on periods of increased cell proliferation. This process depends upon the fidelity and efficiency of DNA replication in order for genetic information to be copied and distributed to new cells in a timely and accurate manner. During normal, uninhibited eukaryotic replication, highly processive error-free polymerases Pol δ

and Pol ε are responsible for synthesizing the nascent DNA strands in a 5' to 3' direction, both continuously along the leading strand (Pol ε) and discontinuously along the lagging strand (Pol δ) of the replication fork (Burgers & Kunkel, 2017). These polymerases belong to a group called the B-family, which also includes the primase Pol α and, due to structural similarities, the trans-lesion synthesis polymerase Pol ζ. In order to lend specificity and processivity to the replication process, the active sites of the main replicative polymerases are smaller and highly specific. However, this same property of error-free replicative polymerases is what tends to prevent them from bypassing lesions that arise during replication. Failure to bypass lesions to the DNA can cause the stalling or collapse of the replication fork. This can then result in the creation of double strand breaks, and during these periods of frequent cell division, the repair machinery necessary to repair these double strand breaks may not be able to keep up. Postponing repair of these lesions in favor of error-prone bypass maintains replication fork progression, and thereby avoids increased the danger for potential collapse. This may actually be the safer option for the cell when the level of DNA damage overwhelms the repair machinery. To avoid the cascade of detrimental events that can occur when the replication machinery is faced with a lesion, cells can employ the DNA Damage Tolerance (DDT) pathway to accomplish bypass.

DDT depends on a suite of specialized DNA polymerases, collectively referred to as trans-lesion synthesis (TLS) polymerases, whose main role is to synthesize past various types of DNA lesions. Unlike the replicative polymerases, the TLS polymerases have low-fidelity, low-processivity, and wider, more

flexible active sites (Vaisman & Woodgate, 2017). These characteristics allow TLS polymerases to accommodate damaged bases in their active sites. Many of them have evolved to recognize certain types of damaged bases, and to incorporate certain bases across from the damage. They also lack the 3' to 5' proofreading activity of processive, high-fidelity polymerases. Their low processivity allows them to dissociate from the template once lesion bypass is complete, after which the normal replisome can be reconstituted, and processive synthesis can continue (Ghosal & Chen, 2013). Having a variety of TLS polymerases at their disposal allows eukaryotic cells to be prepared to survive even extensive levels of damage during replication.

The main TLS polymerases are the Y-family polymerases and the B-family polymerase Pol ζ. Although Pol ζ falls within the family which contains the high-fidelity replicative polymerases Pol δ, Pol ε, and Pol α, as a TLS polymerase it lacks their processivity and proofreading capability (Chun & Jin, 2010). One of the main methods of control thought to initiate recruitment of TLS polymerases in response to damage is the ubiquitination of the proliferating cell nuclear antigen (PCNA) sliding clamp. Within the replisome, PCNA is a trimeric ring-shaped complex which clamps around dsDNA and associates with polymerases such as Pol δ and Pol ε to maintain their processivity as they translocate along with the replication fork (Sale, Lehmann, & Woodgate, 2012). PCNA was first identified as a target of ubiquitination in the context of DDT in yeast (Hoege, Pfander, Moldovan, Pyrowolakis, & Jentsch, 2002; Ulrich, 2009). The same study also demonstrated that PCNA could direct DDT toward either

error-prone or error-free repair by being either be mono-ubiquitinated by the RAD6-RAD18 complex or polyubiquitinated by the UBC13-MMS2-RAD5 complex, respectively. These complexes which direct ubiquitination are each made up of a combination of ubiquitin ligases (RAD18, RAD5) and ubiquitin-conjugating enzymes (RAD 6, UCB13-MMS2) (Hoege et al., 2002). All ubiquitination was observed after the application of sub-lethal levels of DNA damage. Not long after these results were published, two human functional homologs of yeast Rad5 were identified: SNF2 histone-linker PHD RING helicase (SHPRH) and helicase-like transcription factor (HLTF) (Unk et al., 2006, 2008). All of the TLS polymerases have been shown to possess ubiquitin-binding motifs which increased their affinity for ubiquitinated PCNA (Sale et al., 2012). All of this evidence made it clear that this mechanism of DDT control was highly conserved in eukaryotes.

While modification of PCNA seems to be an important component to control of DDT, other proteins are likely to help coordinate the process. In this aspect, REV1 is particularly interesting among its Y-family peers. REV1 is highly conserved in eukaryotes, but the sum of its roles in DDT remains unclear. The polymerase activity of REV1 is limited to inserting dCMP nucleotides across from normal Gs, adducted-Gs, and abasic sites (Waters et al., 2009). However, mounting evidence supports that Rev1's most important roles are in scaffolding, recruitment, and ultimately pathway choice between TLS and TS. Evidence in mammals and other eukaryotes has demonstrated that REV1 is able to interact with all of the other TLS polymerases, specifically through interactions at its CTD

(D'Souza & Walker, 2006; Guo et al., 2003; Kosarek et al., 2008; Tissier et al., 2004). Furthermore, a recent study done in a human U2OS cell line demonstrated that REV1 could stimulate the mono-ubiquitination of PCNA in a mechanism dependent on ubiquitinated RAD18 (Wang et al., 2016). This effect was persistent after DNA damage by UV, mitomycin-C, and hydoxyurea (HU), but not the alkylating agent MMS. In another study in U2OS cells, treatment with MMS led to de-ubiquitination of RAD18. De-ubiquitination of RAD18 had been shown to induce an interaction with SHPRH, potentially in an effort to shunt DDT away from TLS and toward a less error prone pathway—presumably TS (Wang et al., 2016; Zeman, Lin, Freire, & Cimprich, 2014). Together, these studies provided evidence for unique aspects of REV1: first, that it could not only respond to modifications of PCNA, but could also prompt them. Second, that—aside from its role in TLS—Rev1 might play a role in redirecting DDT toward TS in certain contexts.

Recent work has investigated the clinical implications of research on TLS polymerases. Specifically, one study showed that single nucleotide polymorphisms in REV1 and REV3L, the catalytic subunit of Pol ζ, resulted in decreased rates of survival in osteosarcoma patients taking cisplatin-based chemotherapeutics, which target rapidly dividing cancer cells by causing interstrand-crosslink (ICL) damage (Goričar et al., 2015). Survival rates were lower in patients with SNPs in both REV1 and REV3L. This argued for the importance of REV1 as a potential target for research and cancer therapeutics.

How Rev1 mediates or is involved in TS remains to be clarified, and considering that most of the work on Rev1 in DDT has been done in yeast and mammalian cell lines, clarifications in intact multicellular organisms are lacking. To address these gaps, *D. melanogaster* provides an ideal model. Prior work in our lab used a DmRev1 complete deletion mutant and a mutant lacking the CTD to characterize sensitivity to DNA damaging agents (Khodaverdian et al., in preparation). This work demonstrated that *rev1* mutants were hypersensitive to MMS, showing low relative homozygote survival at doses as low as 0.0025% (v/v). This result indicated that DmRev1 was essential to repair of alkylating damage. Surprisingly, a *Rev1-ΔCTD* mutant only became sensitive at a dose of 0.03% MMS (v/v). This level of sensitivity was nearly identical to that of a *rev3* knockout, which eliminates the catalytic subunit of Pol ζ. In yeast, the UV sensitivity phenotypes of REV1 mutants were found to resemble those of REV3 mutants (Lemontt, 1971). The large size of the REV3 protein in mammals had made comparisons of a similar nature difficult in these systems, but knocking out REV3 resulted in embryonic lethality in mice, while REV1 knockouts were viable (Jansen, Tsaalbi-Shtylik, & de Wind, 2015). Our data in *D. melanogaster* indicated a role for DmRev1 outside TLS, independent from its ability to recruit TLS polymerases (**Figure 3-1**). Interestingly, although the same work from our lab also demonstrated that *rev1* mutants were not sensitive to mutagens that induced DSBs, we also showed that a *rev3 brca2* double mutant was as sensitive to MMS as the *rev1* mutant. This suggests that in DDT, when TLS is not an option, homologous recombination is used in an alternative bypass process. This

work laid the foundation for an investigation into this poorly understood branch of DDT.

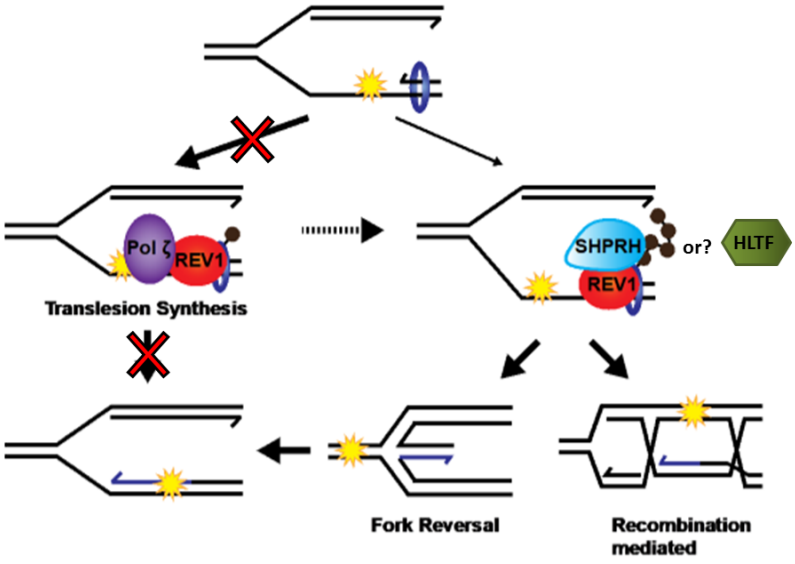


Figure 3-1. Proposed roles of Rev1 in DDT. The pathway above represents DDT occurring after a replication fork encounters a lesion. Bypass of the lesion is accomplished either by TLS, shown on the left, or by TS, shown on the right. When TLS is compromised, we expect DDT to shunt toward TS. The result of a template switch in the TS pathway could be either a fork reversal event or a recombination mediated event. Both can achieve lesion bypass. Our results indicate an essential role for DmRev1 outside of TLS, and independent of its ability to recruit other TLS polymerases via interactions with its C-terminal domain. The TS pathway may depend on the poly-ubiquitination of PCNA (shown as a blue ring with ubiquitin represented by small brown circles), and may also be coordinated by the association of DmRev1 with Rad5-like ubiquitin-ligases, like SHPRH or HLTF.

Modifications on a traditional EMS mutagenesis genetic screen allow specification and faster candidate identification and filtration

In order to discover other uncharacterized proteins which may be involved in DmRev1's role in the TLS-independent and recombination-mediated TS pathway, we employed a modified version of an EMS mutagenesis screen, and focused our analysis on the third chromosome, where the *rev1* gene is located. In *D. melanogaster*, the EMS screen is a classic approach that allows for the unbiased generation of genome wide variation. EMS has been primarily shown to preferentially ethylate the N-7 of guanine, but is capable of creating multiple types of transitions, transversions, and, occasionally, even small deletions (Pastink, Heemskerk, Nivard, van Vliet, & Vogel, 1991; Sega, 1984). In order to target genes potentially involved in the TS pathway of DDT, we made a novel modification to the traditional EMS mutagenesis screen by conducting it in a TLS-deficient background. We accomplished this by using our *Rev1-ΔCTD* stock. By screening for mutants that displayed MMS hypersensitivity phenotypes, we could selectively discover mutations that may have compromised genes in the TS pathway. Examples of potentially affected genes might be the fly homologs of SHPRH or HLTF (**Figure 1**). Recent studies in yeast have shown that Rev1 associates with Rad5 through an interaction at its C-terminal domain (Kuang et al., 2013; Xu et al., 2016). In *D. melanogaster*, we have previously shown that putative yeast Rad5/human SHPRH ortholog *CG7376*, while not sensitive to MMS, became hypersensitive when combined with the *Rev1-ΔCTD* allele in a double mutant (Schmidt, 2016). A functional ortholog for HLTF has not yet been characterized in *D. melanogaster*, although conservation-based orthology predictions identify the gene *lodestar* as a potential ortholog. Finally, addition to

genes potentially involved in TS, our screen might identify genes acting upstream of DDT, which could be involved in signaling.

The screen produced 23 MMS-hypersensitive mutants as candidates, and our next task was to identify the variants causing the sensitivity phenotype in each. The first of these results are presented here. Following the recent example of several other groups (Blumenstiel et al., 2009; Gerhold et al., 2011; Gonzalez et al., 2012; Haelterman et al., 2014; Lee et al., 2016), we used WGS to eliminate the traditional genetic mapping bottleneck that normally impedes the downstream analysis of mutant candidates and the identification of causative variants.

RESULTS

EMS-generated mutants selected for analysis

Twenty-three mutant strains displaying an MMS-hypersensitive phenotype were obtained from an EMS-mutagenesis screen, as described above. The screen was targeted to the third chromosome, and mutants were generated over a *TM3, Ser, P{Act-GFP, w^+}* balancer. Sensitivity of mutants to MMS was determined by measuring the relative percent survival of homozygotes (phenotype: w^-). Relative survival ratios were calculated by comparing the percent survival of homozygotes in a treatment condition (0.01% or 0.03% v/v MMS treatment) to the percent survival of homozygotes in a control condition (water treatment). The dosages tested were based on the pre-determined sensitivity of the *rev1* mutants at 0.01%

MMS and the *Rev1-ΔCTD* mutants at 0.03% MMS, as discussed previously. Given the context of a *Rev1-ΔCTD* background, we predicted that mutants displaying sensitivity at 0.01% MMS (hypersensitivity) might indicate genes compromising the TS pathway.

After hypersensitive mutants were identified, we used complementation testing to determine if any of these mutants were in the same complementation group. Placement of any number of mutants into a complementation group would indicate that these mutants contained separate mutations in the same gene. This is based on the assumption that two mutants with recessive mutations in different genes, when crossed together, should complement. In other words, they should be able to produce viable, transheterozygous progeny that are *not* sensitive to MMS. Alternatively, mutants that do no complement should produce transheterozygous progeny that are sensitive to MMS. Mutants from the latter category were categorized as belonging to a complementation group. While the possibility exists for complementation testing to falsely assign nearby mutations to the same gene, as discussed in Hawley & Gilliland, 2006, we were prepared to account for this experimental assumption with the availability of WGS data for each mutant.

Mutant / cross	**Hom./transhet. relative survival indicated sensitivity to 0.01% MMS?**
Mutant 148	✓ (hypersensitive)
Mutant 157	✓ (hypersensitive)

148*/157 transhet. (complementation test)	✓ (hypersensitive)
157*/148 transhet. (complementation test)	✓ (hypersensitive)
Mutant 1396	✓ (hypersensitive)
1396*/157 transhet. (complementation test)	X (not sensitive)

Table 3-1. MMS sensitivity of mutants analyzed in this study

*Maternally inherited third chromosome genotype listed first

Any mutant strains showing non-complementation upon being crossed were subjected to a reciprocal male/female cross. This allowed us to rule out sex-linked effects in non-complementing crosses. We found two mutants, 148 and 157, which failed to complement and prepared them for WGS (**Table 3-1**). In order to maximize the efficiency of the sequencing run, while still allowing for relatively high coverage per genome, we also chose to include another mutant, 1396, which complements with the 148/157 complementation group. Like 148 and 157, the 1396 mutant also displayed sensitivity to MMS at 0.01% (**Table 3-1**). Additionally, because we do not expect the causative locus in 148/157 to be the same as that in 1396, sequencing all three genomes allowed us to build in another set of background variants that could be removed in downstream analysis (discussed below).

WGS and data analysis pipeline effectively identifies candidate causative variants in non-complementing mutants 148 and 157

Genomic DNA was isolated from either homozygous adult male flies (for isogenized *Rev1-ΔCTD* stock) or from larvae (for mutants) using a phenol-chloroform extraction. The *Rev1-ΔCTD* stock was isogenized before being mutagenized in order to provide a consistent reference genome for downstream sequencing analysis. DNA library preparation from genomic DNA relies on clean, high-quality genomic DNA, and so we validated genomic DNA by gel electrophoresis, Nanodrop, and finally by DNA, RNA, and Protein Qubit. Any genomic preps with contaminating RNA were treated with RNaseA and column purified. We then used the Illumina Nextera kit to create DNA libraries. We checked the library fragment size distribution using an Advanced Analytics Fragment Analyzer. Illumina MiSeq was used to sequence all genomic preps using 2x300 paired-end sequencing runs. Raw sequencing output from WGS was put through a bioinformatics pipeline (pe_align.sh) to generate third chromosome variant calls for each genome in the form of VCF files. Variant calls were made by comparison to the most recent build of the *Drosophila melanogaster* genome: dm6 (Dos Santos et al., 2015). To remove background variants in the mutant genomes that were also present in both the *Rev1-ΔCTD* stock, and would therefore not be responsible for MMS hypersensitivity, we used a program called bedtools. The bedtools software allowed us to intersect or subtract VCF datasets from each other. We then used information about our complementation group to remove additional background present in mutants shown to complement, such that VCF data for 1396 were subtracted from VCF data for 148 and from VCF data for 157. Similarly, VCF data for 148 and 157 were subtracted from VCF data for

1396. The resultant post-bedtools VCF files for all analyses were annotated using the Ebsembl Variant Effect Predictor (VEP). For each data set analyzed, we saved both a full variant list, and a list containing only variants of high or moderate predicted impact. All annotated variant lists from VEP were downloaded and filtered for unique variant changes by simultaneously removing rows with duplicates in all of the following VEP fields: Allele, Symbol, Gene, Amino Acid, and Codon. The resultant VEP-generated variant lists are summarized in **Table 3-2**. On average, only 2.9 % of genes containing variants had variants with a high or moderate predicted impact. For the non-complementing mutants, we directly compared the list of 45 genes in 148 with the list of 53 genes in 157 to find overlaps. This comparison generated a list of just 7 genes that contained moderate to high impact variants in both mutants.

Table 3-2. WGS variant data for mutants analyzed in this study

Data Set Analyzed	Total Variants	Total Genes	Genes with HIGH - MOD. Predicted Impact Variants	Genes with HIGH - MOD. Impact, in list for both non-complementing strains
Mutant 148	3,339	1,591	45	7
Mutant 157	3,962	1,666	53	
Mutant 1396	2,653	1,338	37	N/A

We then went back to the variant list to view the variant calls at these loci in each non-complementing mutant strain. We discovered that only one of these genes, *polybromo*, had unique variants of high predicted impact in 148 and 157,

as well as an additional, unique missense variant in 157 (**Table 3-3**). The variant calls in *polybromo* were validated by viewing the sequencing alignments in the Integrative Genomics Viewer (IGV). Alignments at this locus in 148 and 157 both had high coverage and demonstrated the confidence of the variant calls.

Table 3-3. Variants present in candidate gene *polybromo* in mutants 148 and 157

Gene affected	Variant in 148	Variants in 157
polybromo	HIGH impact – Allele: A (gained stop codon, exon 4/4, Q/STOP amino acid change)	HIGH impact – Allele: T (splice acceptor variant, intron 2/3)
		MODERATE impact – Allele: T (missense variant, exon 2/4, G/E amino acid change)

This analysis identified the variants in *polybromo* as a high priority candidates for validation in mutants 148 and 157. However, other variants in the list of genes affected in 148 and 157 may also need to be considered if the variants in *polybromo* cannot be confirmed as responsible for the MMS hypersensitivity phenotype in these mutants (see **Table A-3-1** in Appendix for a list of the 7 genes containing variants of high or moderate predicted impact in mutants 148 and 157).

WGS and data analysis pipeline effectively identifies several strong candidate genes in single MMS hypersensitive mutant 1396

After demonstrating the utility of this WGS framework to identify a strong candidate for a causal gene in two non-complementing MMS hypersensitive mutants from our EMS mutagenesis screen, we also wanted to analyze the list of genes present in our single mutant, 1396. In this case, we started with a larger list of genes in consideration: 37 total (of which 13 had only high predicted impact variants, 22 had only moderate predicted impact variants, and 2 had both types). We identified two particularly interesting genes within this list, *Claspin* and *lodestar*. Claspin/CLSPN is a nuclear protein involved in the ATR-Chk1 checkpoint, and is highly conserved in multicellular eukaryotes (Gramates et al., 2017). Lodestar is a SNF2 family helicase-like protein which has been shown to be involved in the organization and segregation of chromatin in *D. melanogaster*, but orthology predictions identify it as a possible ortholog of yeast Rad5 and human HLTF (Gramates et al., 2017; Szalontai et al., 2009). Therefore, *Claspin* and *lodestar* stood out due to their potential to influence the DDT pathway, and they contain strong candidates for causative variants in 1396 (**Table 3-4**). The variant calls in *Claspin* and *lodestar* were validated by viewing the sequencing alignments in the Integrative Genomics Viewer (IGV). Alignments at the *Claspin* locus revealed that the relevant variant was actually in an area of very low coverage. At this position, GATK called a T missense variant in 1396, but did not call any variant in the *Rev1-ΔCTD* genome. The VCF files for 1396 and *Rev1-ΔCTD* therefore differed at this position, allowing the variant to come through as a candidate in our bedtools-generated list. However, upon viewing the 1396 and *Rev1-ΔCTD* alignments side by side, it was clear that the few reads present at this

location in the *Rev1-ΔCTD* did actually show the same T variant, despite the fact that it was missed by GATK. These observations indicated that this variant was actually a false positive, and underlined the importance of viewing alignments in choosing causative variant candidates. Alignments at the variant locus of interest in *lodestar* showed high quality coverage, and indicated that the variant call represented a true candidate for a causative variant.

Table 3-4. Variants present in candidate genes *Claspin* and *lodestar* in mutant 1396

Gene affected	Variant in 1396	Confirmed by alignments?
Claspin	MODERATE impact – Allele: T (missense variant, exon 1/7, K/M amino acid change)	NO
lodestar	HIGH impact – Allele: T (gained stop codon, exon 2/5, Q/STOP amino acid change)	YES

This analysis identified the variant in *lodestar* as a high priority candidate for validation in mutant 1396. However, other genes in the list generated for 1396 may also need to be considered if the variant in *lodestar* cannot be confirmed as responsible for the MMS hypersensitivity phenotype in this mutant (see **Table A-3-2** in Appendix for a list of all genes containing variants of high or moderate predicted impact in mutant 1396).

DISCUSSION

The highly conserved DDT pathway is crucial to the maintenance of genomes during periods of development when DNA synthesis demands are high. DDT allows the cells to bypass the lesions that threaten to cause fork collapse and stalling in favor of later repair by mechanisms such a base excision or nucleotide excision repair (BER or NER). While this makes DDT a mutagenic process, it also saves the cell from far more detrimental consequences such as DSBs, chromosomal fragility, and eventually, possible cell death. The DDT pathway consists of two main sub-pathways for bypass: TLS and TS. We have previously shown that one TLS polymerase, Rev1, is essential to both in *D. melanogaster* (**Figure 3-1**). The role of REV1 in the TLS pathway is well characterized, and depends on its ability to recruit other TLS polymerases that specialize in synthesizing across from and past lesions. Previous studies in our lab demonstrated that DmRev1 had a role in DDT that was independent of its ability to recruit TLS polymerases. This implicated Rev1 in TS, and potentially in coordinating DDT pathway choice. We modified a traditional EMS mutagenesis screen by conducting it in an isogenized *Rev1-ΔCTD* background to focus on third chromosome mutations that might compromise the TS pathway. The mutants from the screen showing hypersensitivity to MMS represented mutations that could have compromised genes involved in TS. Here, we have presented the application of a WGS and bioinformatics analysis pipeline to multiple mutants obtained from the screen, and identify strong third chromosome candidate genes with potentially causative variants. We demonstrated that this pipeline is

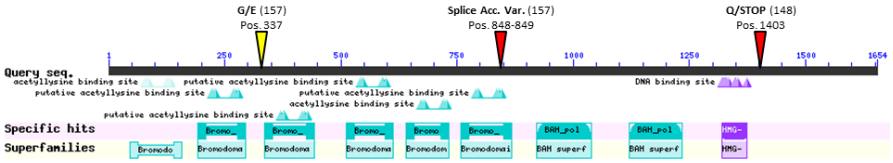
sufficient to generate manageable candidate lists from the two different types of data sets tested. The availability of complementing mutants within a group does seem to increase the power of the analysis, but the list generated for the single mutant was still manageable in size. Most importantly, if validated, either of the two candidates we have presented here will expand our current understanding of mechanisms and control of TS in *D. melanogaster* DDT.

We first applied our analysis to identify candidates for causative variants in a set of non-complementing mutant strains, 148 and 157, and identified *polybromo* as our top candidate (**Table 3-2, 3-3**). This gene is highly conserved in eukaryotes, and in *D. melanogaster*, *polybromo* is thought to be involved in chromatin remodeling as part of the SWI/SNF-like brama complex (Gramates et al., 2017; Vorobyeva, Mazina, & Doronin, 2013). Studies have shown that *polybromo* is important to eggshell development in ovarian follicle cells (Carrera et al., 2008; Gramates et al., 2017). We decided to take the first steps toward validating our variants in *polybromo* as causative by screening for the same eggshell defects in 148 and 157 that were observed in the *D. melanogaster polybromo* mutants from the 2008 study. Preliminary data has not shown eggshell defects in our screen of 148/157 transheterozygous embryos. However, it is still possible that the variants we identified in *polybromo* are responsible for MMS sensitivity in 148 and 157. In other words, the variants in these mutant strains may compromise *polybromo* in DDT, but not in eggshell and chorion development (discussed further below). In either scenario, further validation will be necessary to confirm that *polybromo* is in fact responsible for the MMS hypersensitivity in

148 and 157. For example, if we can express a transgenic copy of *polybromo* in each mutant and rescue MMS hypersensitivity, this will provide further confidence in validation. We have also begun validation of our *polybromo* variants by Sanger sequencing, and thus far have successfully validated both mutants in strain 157.

Evidence supports a connection between chromatin remodeling and DDT. Chromatin remodelers are involved in the ATP-dependent positioning of nucleosomes to allow access to DNA by replication, transcription, and repair machinery (Saha, Wittmeyer, & Cairns, 2006). In addition, they are responsible for the normal progression of DNA replication forks and for their progression during times of replication stress (Vincent, Kwong, & Tsukiyama, 2008). Studies in yeast have implicated chromatin remodelers in modifying PCNA and recruiting repair proteins to sites of DNA damage (Downs et al., 2004; Falbo et al., 2009). The later of these two studies implicated the yeast INO80 chromatin remodeler in DDT specifically. An *ino80* mutant showed hypersensitivity to MMS, increased accumulation of γ-H2AX after progression through S-phase, reduced PCNA ubiquitination, and demonstrated a failure to recruit RAD18. Perhaps their most interesting result was that *ino80* mutants failed to recruit Rad51 to replication forks, a process normally in place to ensure that recombination-mediated repair is available in the case of blocked replication (Falbo et al., 2009). The clinical significance of Polybromo-1 was highlighted by another study which showed that mutations in *PBRM1* were found in 41% (92 out of 227) clear cell renal cell carcinoma ccRCC specimens (Varela et al., 2011). Perhaps the most implications

to our work are underlined by more recent work in human cells, which demonstrated that Polybromo-1 (*PBRM1/BAF180*) played a role in the ubiquitination of PCNA after treatment with both UV and hydroxyurea (HU), and that this particular role was independent of the ATPase activity used to position nucleosomes (Niimi, Hopkins, Downs, & Masutani, 2015). This study also



demonstrated that, after UV damage, a mutant construct containing only the bromo-adjacent homology (BAH) domains was sufficient to induce PCNA ubiquitination, despite the fact that this mutant could not be assembled into PBAF, a SWI/SNF family chromatin remodeling complex. Taken together, these experiments suggest that the PCNA ubiquitination activity of Polybromo-1 is independent of its ATPase-dependent chromatin remodeling activity.

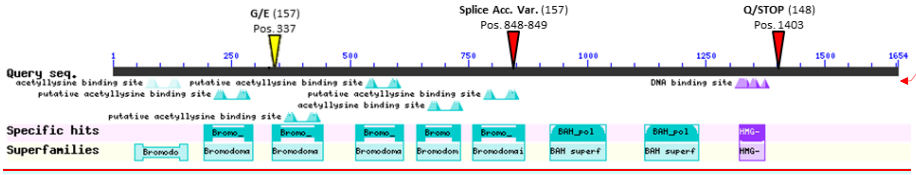


Figure 3-2. Variant locations within DmPolybromo in 148 and 157. The NCBI conserved domain search tool was used to create this figure, which shows conserved bromodomains, Bromo-Adjacent Homology (BAH) domains, and the High Mobility Group (HMG) domain in DmPolybromo. The variants from mutants 148 and 157 are shown by colored arrowheads. Variant

Formatted: Indent: First line: 0"

descriptions are in bold, followed by the strain in parentheses. Protein positions are indicated. Red arrowheads represent variants of high predicted impact. The yellow arrowhead represents a variant of moderate predicted impact.

These findings in human cell lines concerning the BAH domains of Polybromo-1 lend support to the observations we have made in mutants 148 and 157. Because the high impact splice acceptor variant in 157 is positioned immediately upstream of the BAH domains, it is possible that the PCNA ubiquitination activity is removed, but the N-terminal end of the protein, and thus its bromodomains, remain unaffected (**Figure 3-2**). Of course, this would also rely on the assumption that the missense variant also present in strain 157 does not compromise the function of the third bromodomain in which it is located. However, even if it does, it is possible that this change is not detrimental, or that the other bromodomains can compensate. In fact, one recent study in human cell lines demonstrated that the third bromodomain may not be as important to the chromatin remodeling function of Polybromo-1 as the others (Porter & Dykhuizen, 2017). Because the gained stop codon in strain 148 is located at the far C-terminal end of the protein, it is difficult to predict what effect this variant would have on DmPolybromo function (**Figure 3-2**). However, because 148 and 157 fail to complement, it remains possible that this variant has a detrimental effect. Thus far, the most well-studied *polybromo* mutant allele studied in *D. melanogaster* is a complete protein null; female flies with this allele were sterile, producing inviable eggs with irregular chorion, as described previously (Carrera

et al., 2008; He et al., 2014). Taken together, this argues for the variants we have isolated in *polybromo* to be very functionally different alleles of the gene.

Identifying *polybromo* as an interacting protein in the control or mechanism of TS in *D. melanogaster* would provide novel insight. The combined information from work in human cell lines and in flies argues for our variants as alleles of *polybromo* that may only affect its role in PCNA ubiquitination. Extensive work has demonstrated that PCNA ubiquitination is central to DDT pathway choice. It is possible that in *D. melanogaster* replication, *polybromo* could be essential to ubiquitination of PCNA after MMS induced damage. If *polybromo* can initiate poly-ubiquitination of PCNA specifically, this would allow for lesion bypass by a Rad51-mediated TS type of mechanism in a TLS-deficient context (*Rev1-ΔCTD* mutant background).

Comment [SBT11]: Added and revised here, plus the new figure

The same analysis applied to non-complementing mutants 148 and 157 was also applied to a single MMS hypersensitive mutant: 1396. While the final list of genes under consideration in this data set was larger, one high priority candidate for a causative variant in the lodestar gene was identified (**Table 3-3**). *Lodestar* contained a gained stop codon variant of high predicted impact in mutant 1396 (**Table 3-4**). Further phenotypic and genetic validation will be necessary to determine the ultimate strength of the variant in lodestar as the best causative candidate in mutant 1396, but this gene is particularly interesting, due to its potential ties to DDT. In eukaryotes lodestar is highly conserved, and has been implicated as an ortholog to yeast Rad5 and mammalian HLTF. Furthermore, the

design of our mutagenesis screen would clearly implicate lodestar in the tolerance of MMS damage in the absence of TLS, if its variant can be confirmed as causative. Taken together, this would argue strongly for lodestar as a key component of the TS pathway of DDT in D. melanogaster. Thus far, lodestar has been identified as a helicase-like protein in D. melanogaster, and a dominant negative mutation was shown to have abnormal chromosome segregation, increased nondisjunction, and female sterility (Szalontai et al., 2009). Orthology predictions for lodestar collected on Flybase suggest its possible tie to human HLTF (Gramates et al., 2017), but also lists the primary function of DmLodestar as transcription termination.

Figure 3-3. *D. melanogaster lodestar* is a strong candidate for a HLTF ortholog. (a) Domain predictions in *D. melanogaster lodestar* obtained from protein BLAST (pBLAST). (b) Alignment scores from a two-sequence alignment pBLAST between human HLTF and *D. melanogaster lodestar*, which shows the high level of identity (alignment score ≥ 200, identity = 31%) at the C-terminus, which contains the predicted ATPase domain in *lodestar*. (c) Detailed view: CLUSTALW alignment in the C-terminal region of *D. melanogaster* lodestar and human HLTF. (***** = fully conserved residue; **:** = strongly similar properties; **.** = weakly similar properties)

To investigate this further, we used CLUSTALW (2.1) to align human HLTF and DmLodestar (**Figure 3-3**). Near the putative ATPase region, these proteins share 31% identity. Pending the confirmation of the stop codon variant in *lodestar* being causative for the MMS hypersensitivity in strain 1396, the evidence we have collected thus far would strongly *lodestar* as the possible *D.*

melanogaster ortholog of HLTF. In human fibroblasts, knockdown of HLTF has been shown to lead to sensitivity to both UV and MMS (Unk et al., 2008). However, a later study argued for damage specific roles for HLTF and its other Rad5 ortholog, SHPRH (Lin, Zeman, Chen, Yee, & Cimprich, 2011). In this study, MMS treatment resulted in ubiquitination and degradation of HLTF, and increased mutagenesis in an SHPRH-deficient background, but not an HLTF-deficient background. An HLTF-deficient background instead resulted in increased UV-induced mutagenesis. It is possible that, despite key functional domain similarities in *lodestar* and HLTF between flies and humans, damage specificity may not have been conserved.

Our *lodestar* variant in mutant 1396 is a gained stop codon occurring early in the protein, upstream of both the SNF2_N superfamily and HELICc domains (**Figure 3-4**). Characterized alleles of *lodestar* in *D. melanogaster* have demonstrated strong maternal effect lethality, chromosomal defects, and in the case of the Horka^{D} allele (Ala777Thr), a dominant negative effect which may be caused by an increased affinity for chromatin (Erdélyi & Szabad, 1989; Girdham & Glover, 1991; Szabad, Mathe, & Puro, 1995; Szalontai et al., 2009). Interestingly, the 1396 mutant produced viable, albeit MMS-hypersensitive, homozygotes. Based on the previous results in *D. melanogaster*, we might expect that an upstream stop codon in *lodestar* would not produce viable homozygotes. Therefore, if we can confirm this mutation by Sanger sequencing, it will be interesting to follow up on characterization and validation of this variant. A further complication exists in the variant-proximal location of a small Cajal-body

specific RNA (scaRNA:MeU2-C41). The transcribed sequence of this scaRNA begins 2 bp downstream of the variant of interest in 1396. While it is difficult to predict on the impact of this variant on the scaRNA, it is important to consider during efforts to validate the gained stop codon in *lodestar* as causative in MMS hypersensitivity of mutant 1396.

Comment [SBT12]: Added

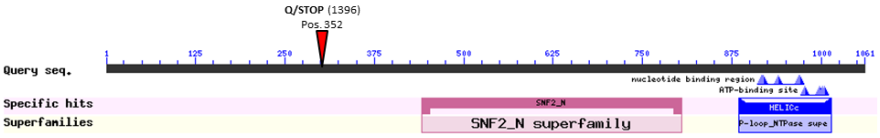


Figure 3-4. Variant locations within DmLodestar for 1396. The NCBI conserved domain search tool was used to create this figure, which shows the conserved SNF2 family N-terminal domain, and the conserved helicase c-terminal (HELICc) domain in DmLodestar. The variant from mutant 1396 is shown by a colored arrowhead (red color indicates that the variant is of high predicted impact). Variant description is in bold, followed by the strain in parentheses. Protein position is indicated.

Final confirmation of causative variants responsible for the MMS sensitivity phenotype in mutants 148, 157, and 1396 will require further validation, but what we have presented here demonstrates the utility of a powerful WGS and bioinformatics pipeline for filtration and prioritization of candidates for causative variants in a modified EMS mutagenesis screen.

MATERIALS AND METHODS

Genomic DNA Extraction from whole flies and larvae

All strains were generated over the *TM3, Ser, P{Act-GFP, w^+}* third chromosome balancer, so that homozygous adults and larvae could be easily identified. For MMS-hypersensitive mutants analyzed in this study, heterozygous siblings were crossed and allowed to lay eggs on grape-juice agar plates for 24 hours. 50 GFP- late second instar and third instar larvae were collected from grape-juice agar plates and rinsed with deionized water, then frozen down at -80°C. For the *Rev1-ΔCTD* stock, 30 homozygous male adults were collected and frozen down at -80°C. Larvae or flies were homogenized in 600 µL DNA extraction lysis buffer (10mM Tris-HCl pH 8.0, 10 mM EDTA, 150 mM NaCl; plus addition of 30 µL of 10 mg/mL Proteinase K, 60 µL of 10% SDS, and 20 µL of 10 mg/mL RNaseA) and incubated at 65°C for 1 hour and allowed to cool to room temperature. Phenol:chloroform:isoamyl alcohol (25:24:1) was added to the lysis mixture (1:1), left on a rocking platform for 15 minutes, and spun down for 15 minutes at 14,000 at room temperature. This was repeated for one additional extraction with phenol:chloroform:isoamyl alcohol, after which a $1/5^{th}$ volume of 8M potassium acetate was added. This was followed by one final extraction with chloroform. DNA was precipitated from the supernatant with isopropanol, and incubated at -80°C for 30 minutes. The pellet was rinsed with 70% ethanol and allowed to dry for 5-10 minutes. Pellets were re-suspended in 50 µL pico pure water. Initial quality control (QC) of genomic DNA was assessed by Nanodrop

spectrophotometry and gel electrophoresis. For increased accuracy, DNA, RNA, and Protein quantification was assed using Qubit spectrophotometry. Residual RNA was removed by treatment with 10 mg/mL RNaseA followed by incubation at 37°C for at least 30 minutes, followed by clean-up using a Macherey-Nagel NucleoSpin Gel and PCR Clean-up kit.

Library Preparation and Illumina MiSeq WGS

Genomic DNA passing QC was diluted to 2.5 ng/µL in 20 µL. DNA libraries for MiSeq WGS were prepared according to the Illumina Nextera protocol for 2x250 paired-end sequencing runs (Illumina Inc., 2016). The Advanced Analytical (AATI) Fragment Analyzer (DNF-474 High Sensitivity NGS Fragment Analysis Kit) was used to assess fragment size distribution of DNA libraries before sequencing (Advanced Analytical Technologies, 2017). Paired-end (2x300) sequencing was done on an Illumina MiSeq. Most runs were confined to 3-4 indexed genomes per flowcell lane in order to balance coverage and efficiency. All index (index adapters) sequences used were from the Nextera kit, and sequence information can be found on the "Index Sequences" page of the Nextera Library Prep Reference Guide (Illumina Inc., 2016).

WGS Data Analysis and filtration

Analysis via the Tufts University High Performance Compute Cluster (HPC): All WGS raw data was provided in the zipped (.gz) file format and transferred to the Tufts HPC using a free, open source file transfer program, FileZilla (https://filezilla-project.org/). We connected to the HPC remotely using another free, open source program: a SSH and Telnet client called PuTTY (http://www.putty.org/). The data manipulations described here were all accomplished via command line (Unix/Linux) and software modules available on the HPC network. WGS raw read .gz files were loaded onto our lab's personal HPC directory, and decompressed to the FASTQ file format. If applicable, FASTQ files for genomes with multiple runs' worth of data (either from multiple full WGS runs, or from a WGS run and a QC run) were concatenated. For each genome, FASTQ files for paired end reads (format: genome1_R1.fastq, genome1_R2.fastq) were processed via a paired-end alignment script (pe_align.sh) designed to generate variant call format (VCF) data. The script employed the following software for data manipulation, in order: Bowtie2 (alignments), SAMtools (quality filtration; SAM to binary format, BAM; sorting), Picard (remove PCR duplicates), SAMtools (index BAM files), GATK (create VCF from BAM). Finally, the script also selectively filtered final VCF output to the third chromosome, and generated VCF files for the left arm, right arm, and entirety of that chromosome (format: genome1_3L.vcf, genome1_3R.vcf, genome1.vcf). The dm6 *Drosophila melanogaster* genome build was used as the reference genome for variant calling in the script. For generation of variant lists, the full chromosome VCF files were compared to each other using bedtools

(version 2.26.0), which was included as a software module within the HPC. To overlap genomes (find similarities), we used bedtools-intersect. To subtract genomes (remove background), we used bedtools-subtract.

Post HPC Analyses and Filtration: After bedtools analysis, VCF files were downloaded via FileZilla, and were then annotated with the Ensembl Variant Effect Predictor (VEP) (http://www.ensembl.org/Tools/VEP) for *D. melanogaster* (BDGP6 genome assembly). Settings used were as follows (any settings not mentioned were not selected). Identifiers: Gene symbol, CCDS, Protein, Uniprot, HGVS, $CSN^{(p)}$. Frequency data (find co-located known variants): Yes. Miscellaneous: Transcript biotype, Protein domains, Exon and intron numbers, Transcript support level, APPRIS, Identify canonical transcripts, Upstream/Downstream distance (bp) – 5000 bp. After obtaining results from VEP, annotated variant information was downloaded as a TXT file. Within VEP, we also filtered results down to a list of variants with high and moderate predicted impact, and downloaded this information as a separate TXT file. To visualize variant lists, TXT files were viewed, manipulated, and saved as MS Excel workbooks. Any exact duplicate rows were removed from these files, Subsequently Allele, Gene, Symbol, Amino Acid, and Codon were removed in order to create a list of unique variant hits. From this data, we generated gene lists from which to seek out interesting candidates for causative variants. For non-complementing strains, we directly compared gene lists to find overlapping gene hits, and then assessed the variants in those genes. Quality of variant calls was

assessed by viewing the duplicate-filtered BAM files (genome1.nodup.bam) files in IGV.

ACKNOWLEDGEMENTS

I would like to thank Mitch McVey for experimental guidance, scientific insight, and for helpful reading and comments for this manuscript. This project was based on the mutants obtained in an EMS-mutagenesis screen initiated by Varandt Khodeverdian, and I thank him for his tireless work and dedication in making this project a reality. Thanks are also due to Tokio Sano, Natalie Danziger, Hannah Slutsky, Sarah Shnayder, Jake Cosgrove, and all the McVey lab members who helped to obtain, characterize, and complement-test the mutants from this screen. Finally, thank you to Reazur Rahman for providing the pe_align.sh script, and to David LaPointe for bioinformatics training and help. This research was supported by grant P01GM105473 from the NIH.

REFERENCES

Advanced Analytical Technologies. (2017). Fragment Analyzer Automated CE System: Quick Start Guide - 12 Capillary.

Blumenstiel, J. P., Noll, A. C., Griffiths, J. A., Perera, A. G., Walton, K. N., Gilliland, W. D., Hawley, R S., Staehling-Hampton, K. (2009). Identification of EMS-induced mutations in Drosophila melanogaster by whole-genome sequencing. *Genetics*, *182*(1), 25–32.

Burgers, P. M. J., & Kunkel, T. A. (2017). Eukaryotic DNA Replication Fork. *Annual Review of Biochemistry*, *86*, 417–438.

Carrera, I., Zavadil, J., & Treisman, J. E. (2008). Two subunits specific to the PBAP chromatin remodeling complex have distinct and redundant functions during drosophila development. *Molecular and Cellular Biology*, *28*(17), 5238–5250.

Chun, A. C. S. & Jin, D. (2010). Ubiquitin-dependent regulation of translesion polymerases. *Biochemical Society Transactions*, *38*(Pt 1), 110–115.

D'Souza, S., & Walker, G. C. (2006). Novel role for the C terminus of Saccharomyces cerevisiae Rev1 in mediating protein-protein interactions. *Molecular and Cellular Biology*, *26*(21), 8173–8182.

Dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., Emmert, D. B., Gelbart, W. M., Brown, N. H., Kaufman, T., Werner-Washburne, M., Cripps, R., Broll, K., Gramates, L. S., Falls, K., Matthews, B. B., Russo, S., Zhou, P., Zytkovicz, M., Adryan, B., Attrill, H., Costa, M., Marygold, S., McQuilton, P., Millburn, G., Ponting, L., Stefancsik, R., Tweedie, S., Grumbling, G. (2015). FlyBase: Introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, *43*(D1), D690–D697.

Downs, J. A, Allard, S, Jobin-Robitaille, O., Javaheri, A., Auger, A., Bouchard, N., Kron, S. J., Jackson, S. P., Côté, J. (2004). Binding of Chromatin-Modifying Activities to Phosphorylated Histone H2A at DNA Damage Sites. *Molecular Cell*, *16*(6), 979–990.

Erdélyi, M., & Szabad, J. (1989). Isolation and characterization of dominant female sterile mutations of Drosophila melanogaster. I. Mutations on the third chromosome. *Genetics*, *122*(1), 111–27.

Falbo, K. B., Alabert, C., Katou, Y., Wu, S., Han, J., Wehr, T., Xiao, J., He, X., Zhang, Z., Shi, Y., Shirahige, K., Pasero, P., Shen, X. (2009). Involvement of

a chromatin remodeling complex in damage tolerance during DNA replication. *Nature Structural & Molecular Biology*, *16*(11), 1167–1172.

Gerhold, A. R., Richter, D. J., Yu, A. S., & Hariharan, I. K. (2011). Identification and characterization of genes required for compensatory growth in Drosophila. *Genetics*, *189*(4), 1309–1326.

Ghosal, G., & Chen, J. (2013). DNA damage tolerance: a double-edged sword guarding the genome. *Translational Cancer Research*, *2*(3), 107–129.

Girdham, C. H., & Glover, D. M. (1991). Chromosome tangling and breakage at anaphase result from mutations in lodestar, a Drosophila gene encoding a putative nucleoside triphosphate-binding protein. *Genes and Development*, *5*(10), 1786–1799.

Gonzalez, M., Van Booven, D., Hulme, W., Ulloa, R., Lebrigio, R., Osterloh, J., Logan, M., Freeman, M., Zuchner, S. (2012). Whole Genome Sequencing and a New Bioinformatics Platform Allow for Rapid Gene Identification in D. melanogaster EMS Screens. *Biology*, *1*(3), 766–777.

Goričar, K., Kovač, V., Jazbec, J., Zakotnik, B., Lamovec, J., & Dolžan, V. (2015). Translesion polymerase genes polymorphisms and haplotypes influence survival of osteosarcoma patients. *Omics: A Journal of Integrative Biology*, *19*(3), 180–185.

Gramates, L. S., Marygold, S. J., Dos Santos, G., Urbano, J. M., Antonazzo, G., Matthews, B. B., Rey, A. J., Tabone, C. J., Crosby, M. A., Emmert, D. B., Falls, K., Goodman, J. L., Hu, Y., Ponting, L., Schroeder, A. J., Strelets, V. B., Thurmond, J., Zhou, P., Perrimon, N., Gelbart, S. R., Extavour, C., Broll, K., Zytkovicz, M., Brown, N. H., Attrill, H., Costa, M., Fexova, S., Jones, T., Larkin, A., Millburn, G. H., Staudt, N., Kaufman, T., Grumbling, G. B., Cripps, R., Werner-Washburne, M., Baker, P. (2017). FlyBase at 25: Looking to the future. *Nucleic Acids Research*, *45*(D1), D663–D671.

Guo, C., Fischhaber, P. L., Luk-Paszyc, M. J., Masuda, Y., Zhou, J., Kamiya, K., Kisker, C., Friedberg, E. C. (2003). Mouse Rev1 protein interacts with multiple DNA polymerases involved in translesion DNA synthesis. *EMBO Journal*, *22*(24), 6621–6630.

Haelterman, N. A., Jiang, L., Li, Y., Bayat, V., Sandoval, H., Ugur, B., Tan, K. L., Zhang, K., Bei, D., Xiong, B., Charng, W. L., Busby, T., Jawaid, A., David, G., Jaiswal, M., Venken, K. J. T., Yamamoto, S., Chen, R., Bellen, H. J. (2014). Large-scale identification of chemically induced mutations in Drosophila melanogaster. *Genome Research*, *24*(10), 1707–1718.

Hawley, R. S., & Gilliland, W. D. (2006, September). Sometimes the result is not the answer: The truths and the lies that come from using the complementation test. *Genetics*, *174*(1), 5–15.

He, J., Xuan, T., Xin, T., An, H., Wang, J., Zhao, G., & Li, M. (2014). Evidence for chromatin-remodeling complex PBAP-controlled maintenance of the Drosophila ovarian germline stem cells. *PLoS ONE*, *9*(7), e103473.

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

Illumina Inc. (2016). Nextera Library Prep Reference Guide. Sample Preparation Guide, (January), 1–28. https://doi.org/RS-122-2001; RS-122-2002

Jansen, J. G., Tsaalbi-Shtylik, A., & de Wind, N. (2015). Roles of mutagenic translesion synthesis in mammalian genome stability, health and disease. *DNA Repair*, *29*, 56–64.

Kosarek, J. N., Woodruff, R. V, Rivera-Begeman, A., Guo, C., D'Souza, S., Koonin, E. V, Walker, G. C., Friedberg, E. C. (2008). Comparative analysis of in vivo interactions between Rev1 protein and other Y-family DNA polymerases in animals and yeasts. *DNA Repair*, *7*(3), 439–451.

Kuang, L., Kou, H., Xie, Z., Zhou, Y., Feng, X., Wang, L., & Wang, Z. (2013). A non-catalytic function of Rev1 in translesion DNA synthesis and mutagenesis is mediated by its stable interaction with Rad5. *DNA Repair*, *12*(1), 27–37.

Lee, C.-H., Rimesso, G., Reynolds, D. M., Cai, J., & Baker, N. E. (2016). Whole-Genome Sequencing and iPLEX MassARRAY Genotyping Map an EMS-induced Mutation Affecting Cell Competition in Drosophila melanogaster. *G3: Genes|Genomes|Genetics*, *6*(10), 3207–3217.

Lemontt, J. F. (1971). Mutants of yeast defective in mutation induced by ultraviolet light. *Genetics*, *68*(1), 21–33.

Lin, J. R., Zeman, M. K., Chen, J. Y., Yee, M. C., & Cimprich, K. A. (2011). SHPRH and HLTF Act in a Damage-Specific Manner to Coordinate Different Forms of Postreplication Repair and Prevent Mutagenesis. *Molecular Cell*, *42*(2), 237–249.

Niimi, A., Hopkins, S. R., Downs, J. A., & Masutani, C. (2015). The BAH domain of BAF180 is required for PCNA ubiquitination. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, *779*, 16–23.

Pastink, A., Heemskerk, E., Nivard, M. J. M., van Vliet, C. J., & Vogel, E. W. (1991). Mutational specificity of ethyl methanesulfonate in excision-repair-proficient and -deficient strains of Drosophila melanogaster. *Molecular & General Genetics*, *229*(2), 213–218.

Porter, E. G., & Dykhuizen, E. C. (2017). Individual bromodomains of Polybromo-1 contribute to chromatin association and tumor suppression in clear cell renal carcinoma. *Journal of Biological Chemistry*, *292*(7), 2601–2610.

Saha, A., Wittmeyer, J., & Cairns, B. R. (2006). Chromatin remodelling: the industrial revolution of DNA around histones. *Nature Reviews Molecular Cell Biology*, *7*(6), 437–447.

Sale, J. E., Lehmann, A. R., & Woodgate, R. (2012). Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nature Reviews Molecular Cell Biology*, *13*(3), 141–152.

Schmidt, A. (2016). On the role of SHPRH and REV1 in DNA Damage Tolerance in Drosophila Melanogaster. Tufts Digital Library. Tufts University.

Sega, G. A. (1984, September 1). A review of the genetic effects of ethyl methanesulfonate. *Mutation Research - Reviews in Genetic Toxicology*, *134*(2-3), 113–142.

Szabad, J., Mathe, E., & Puro, J. (1995). Horka, a dominant mutation of Drosophila, induces nondisjunction and, through paternal effect, chromosome loss and genetic mosaics. *Genetics*, *139*(4), 1585–1599.

Szalontai, T., Gaspar, I., Belecz, I., Kerekes, I., Erdelyi, M., Boros, I., & Szabad, J. (2009). HorkaD, a chromosome instability-causing mutation in drosophila, is a dominant-negative allele of lodestar. *Genetics*, *181*(2), 367–377.

Tissier, A., Kannouche, P., Reck, M. P., Lehmann, A. R., Fuchs, R. P. P., & Cordonnier, A. (2004). Co-localization in replication foci and interaction of human Y-family members, DNA polymerase polη and REVl protein. *DNA Repair*, *3*(11), 1503–1514.

Ulrich, H. D. (2009). Regulating post-translational modifications of the eukaryotic replication clamp PCNA. *DNA Repair*, *8*(4), 461–469.

Unk, I., Hajdú, I., Fátyol, K., Hurwitz, J., Yoon, J.-H., Prakash, L., Prakash, S., Haracska, L. (2008). Human HLTF functions as a ubiquitin ligase for proliferating cell nuclear antigen polyubiquitination. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(10), 3768–3773.

Unk, I., Hajdú, I., Fátyol, K., Szakál, B., Blastyák, A., Bermudez, V., Hurwitz, J., Prakash, L., Prakash, S., Haracska, L. (2006). Human SHPRH is a ubiquitin ligase for Mms2-Ubc13-dependent polyubiquitylation of proliferating cell nuclear antigen. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(48), 18107–18112.

Vaisman, A., & Woodgate, R. (2017). Translesion DNA polymerases in eukaryotes: what makes them tick? *Critical Reviews in Biochemistry and Molecular Biology*, *52*(3), 274–303.

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C. K., Stephens, P., Davies, H., Jones, D., Lin, M. L., Teague, J., Bignell, G., Butler, A., Cho, J., Dalgliesh, G. L., Galappaththige, D., Greenman, C., Hardy, C., Jia, M., Latimer, C., Lau, K. W., Marshall, J., McLaren, S., Menzies, A., Mudie, L., Stebbings, L., Largaespada, D. A., Wessels, L. F., Richard, S., Kahnoski, R. J., Anema, J., Tuveson, D. A., Perez-Mancera, P. A., Mustonen, V., Fischer, A., Adams, D. J., Rust, A., Chan-on, W., Subimerb, C., Dykema, K., Furge, K., Campbell, P. J., Teh, B. T., Stratton, M. R., Futreal, P. A. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*, *469*(7331), 539–542.

Vincent, J. A., Kwong, T. J., & Tsukiyama, T. (2008). ATP-dependent chromatin remodeling shapes the DNA replication landscape. *Nature Structural & Molecular Biology*, *15*(5), 477–484.

Vorobyeva, N. E., Mazina, M. U., & Doronin, S. A. (2013). SWI/SNF Chromatin Remodeling Complex Involved in RNA Polymerase II Elongation Process in *Drosophila melanogaster*. Chromatin Remodelling, 59–76.

Wang, Z., Huang, M., Ma, X., Li, H., Tang, T., & Guo, C. (2016). REV1 promotes PCNA monoubiquitination through interacting with ubiquitinated RAD18. *Journal of Cell Science*, *129*(6), 1223–1233.

Waters, L. S., Minesinger, B. K., Wiltrout, M. E., D'Souza, S., Woodruff, R. V, & Walker, G. C. (2009). Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. *Microbiology and Molecular Biology Reviews*, *73*(1), 134–154.

Xu, X., Lin, A., Zhou, C., Blackwell, S. R., Zhang, Y., Wang, Z., Feng, Q., Guan, R., Hanna, M. D., Chen, Z., Xiao, W. (2016). Involvement of budding yeast Rad5 in translesion DNA synthesis through physical interaction with Rev1. *Nucleic Acids Research*, *44*(11), 5231–5245.

Zeman, M. K., Lin, J. R., Freire, R., & Cimprich, K. A. (2014). DNA damage-specific deubiquitination regulates Rad18 functions to suppress mutagenesis. *Journal of Cell Biology*, *206*(2), 183–197.

APPENDIX

Consideration of genes with variants of moderate predicted-impact

Although high-priority variant candidates for validation were identified for the sequenced mutants discussed in Chapter 3 (variants in *polybromo* for 148 and 157, and a variant in *lodestar* for 1396), full validation of these candidates has yet to occur for this project. In consideration of this, the expanding gene lists are presented below. Before any of the variants in genes in these expanded lists are considered, sequencing alignments must also be viewed to filter out low quality calls and false positives, as discussed previously. **Table A-3-1** contains genes containing high or moderate predicted impacts in non-complementing mutants 148 and 157. **Table A-3-2** contains all genes with variants of high or moderate predicted impact in mutant 1396.

Table A-3-1. List of genes with high or moderate impact variants in mutants 148 and 157

Gene symbol
Ank2
CG15021
CG33213
Myo61F
*polybromo**
prc
ptip

Cells shown in orange represent genes containing variants of high predicted impact in one or both mutants. Some genes containing high impact variants also contain moderate impact variants.
*****Variants in *polybromo* have already been confirmed by sequencing alignments in 148 and 157.

Table A-3-2. List of genes with high or moderate impact variants in mutant 1396

Gene symbol
Acp76A
Ank2
atk
CG3744
CG3984
CG6283
CG7720
CG7852
CG9801
CG10103
CG12163
CG12413
CG13046
CG14692
CG17249
CG17514
CG17698
CG31036
*Claspin**
Clbn
cno
hipk
iPLA2-VIA
Ir92a
lds (lodestar)†
Mpc1
mTerf5
Muc68Ca
Obp73a
Prp3
ptip
Ptp69D
Rint1
side
sle
sls
Sox100B

Cells shown in orange represent genes containing variants of high predicted impact. Some genes containing high impact variants also contain moderate impact variants. *Variant in *Claspin* has already been deemed a false positive. †Variant in *lodestar* has already been confirmed by sequencing alignments.

Chapter 4: Conclusions

Conclusions and future directions

The work discussed here addresses two distinct questions about pathways for DNA repair and genomic integrity. In both contexts, many interactions, mechanisms, and methods of control still require further elucidation. We attempted to address some of these less-well understood aspects of DNA damage response, first by studying the mechanisms of repair and stability achieved by helicases DmBlm and DmHelQ, and second, in the context of DNA damage tolerance by DmRev1. In attempting to answer questions like these through model systems research, scientists encounter significant challenges in trying to isolate and study specific pieces of this very large molecular puzzle. These challenges are a testament to the complex, yet ever well-provisioned network of interactions that stock the toolbox of the DNA damage response. Like so many DNA repair genes, BLM, HELQ, and REV1 are all implicated in the development and progression of various cancers. Enhancements in our understanding of the hierarchies and interconnections of the pathways in which these genes act could allow for the development of more effective and specific cancer therapeutics. On a broader scale, novel discoveries about these highly conserved genes will clarify their roles in ensuring the faithful replication and repair of genomes.

One technology that has expedited such discoveries within the field of DNA repair, and in the field of biological research in general, is DNA sequencing. Specifically, NGS and WGS were revolutionary developments that provided major improvements in both the depth and efficiency of studies in forward and reverse genetics. These were discussed in Chapter 1. The application of WGS and a downstream framework for bioinformatics data analysis to two different

questions in forward genetics was the underlying endeavor that made the work discussed in both Chapters 2 and Chapter 3 possible (**Figure 1-2**). As discussed in Chapter 1, one of the most exciting components of forward genetics studies in *D. melanogaster* and many other model systems—the discovery and confirmation of a causative mutation—was also, traditionally, the most frustrating and labor-intensive. The time and resources necessary to conduct back-crosses to wild-type stocks in order to obtain recombinants for traditional meiotic mapping, or to cross mutants to marker or deletion library stocks, was near prohibitive. For decades, variant discovery was a substantial barrier in the ability of these studies to produce important, meaningful results in a reasonable amount of time. Now, WGS had broken down that barrier, and model system research has already begun to reap the benefits (Table 1-1). In particular, the search for spontaneous mutations, which do not provide the predetermined mutational signature characteristic of certain mutagens, is now faster and more realistic than ever with the application of WGS. By carefully choosing experimental designs and bioinformatics tools for data analysis, WGS can be applied in a productive manner to both the challenging search for spontaneous causative variants, as well as to the more traditional search for causative variants generated in mutagenesis screens.

In Chapter 2, I used WGS to identify candidates for a synthetic lethality phenotype that we identified in *blm* (DmBlm) *mus301* (DmHelQ) double mutants in *D. melanogaster*. I used two well-characterized null alleles of these third chromosome helicase genes to generate double mutants: blm^{N1} and $mus301^{288A}$.

However, the synthetic lethality was not present in all *blm*^{N1} *mus301*^{288A} double mutants. Most double mutant strains were synthetically homozygous lethal at the third instar larval stage, but a few double mutant strains produced viable homozygous adults. Therefore, I hypothesized that this phenotype did not, in fact, represent a synthetic lethality between these two genes alone, but was actually caused by another mutation in a third gene, which had crossed over with the *blm*^{N1} allele and caused larval lethality. My evidence for this specific interaction was gathered by using a visual method of recombination mapping which is referred to here as VCF recombination mapping, or VRM (**Figure 1-3, Figure 2-5a**). I achieved this by using the Integrative Genomics Viewer (IGV) program for direct visualization of VCF data, and to my knowledge, this particular method for simultaneous mapping and visualization of crossovers represents a novel approach. VRM initially supplied a defined 20 Mbp region of interest (ROI) containing candidates for causative events, which was later refined by a crossover in a newer double mutant strain (*bm-2-119*) which moved the left arm ROI boundary from 70B2 to 78C9 (**Figure 2-5b**). By combining VRM with our bioinformatics pipeline, I was able to produce a relatively manageable list of candidate variants from a total of just 11 recombinant strains. Considering the small number of recombinants, this represents a positive step toward a powerful, adjustable, and broadly applicable method of causative variant identification in a forward genetics context.

With a narrowed list of genes, validation efforts can begin (**Table 2-2, Table A-2-1**). Genes containing candidate causative variants can be validated as

causative in a number of ways. A transgenic copy of the gene could be expressed in lethal double mutant strains to see if larval lethality is rescued. Alternatively, if larval lethality is induced by knocking out the gene in a viable double strain using CRISPR-Cas9 genome editing, this would also provide evidence for the gene's causality in the phenotype. If stocks are available with deletions of the candidate gene, or inserted transposable elements, these can be crossed with viable double mutants and screen for restoration of larval lethality. If these validation techniques are successful for a given candidate, the interaction can be tested further by repeating the latter two validation approaches in a *mus301*^{288A} single mutant background. This will answer whether the relationship observed here is a function of the interaction of this gene with *blm* and *mus301*, or an interaction with *mus301* alone. Both possibilities hold the potential to uncover novel interactions of DmHelQ in DNA repair. It would also be worthwhile to dissect imaginal wing discs from third instar larvae and screen for γH2AV foci in the lethal and viable double mutants, and compare this to the level in single mutant stocks. This would indicate the level of DNA breakage in the various mutants, and would help in further characterization of the larval lethality phenotype.

Comment [SBT13]: Added

I took the same WGS and bioinformatics data analysis pipeline that was built for the work in Chapter 2 and adjusted it so that it could be applied to another project in our lab. In Chapter 3, I discussed the background that prompted this project on DmRev1 and the template switching (TS) pathway of DNA damage tolerance (DDT), and how my framework was able to effectively produce causative variants in this new context. Our lab had previously conducted an EMS

mutagenesis screen in a *Rev1-ΔCTD* background, and obtained mutant strains from this screen that demonstrated hypersensitivity to MMS. *Rev1-ΔCTD* mutants are unable to recruit trans-lesion synthesis (TLS) polymerases to achieve bypass at the site of DNA lesions, effectively eliminating this pathway of DDT. Because the screen was conducted in a trans-lesion synthesis deficient background, we hypothesized that some MMS hypersensitive mutants would represent mutational events in which a gene involved in TS was compromised, thereby completely eliminating DDT. In order to efficiently screen through these mutants for causative variants fitting this description, we decided to use WGS. In doing so, we were able to optimize a part of the traditional genetic screen that—before WGS became accessible to research labs—presented a serious barrier in the discovery of causative variants. The WGS framework from the BLM HELQ project proved to be just as robust in this system, if not more so. The improved resolution attainable in this second application of the framework may be due to the availability of a single, isogenized background stock from which all tested mutants originated, and, potentially, to less overall sequence divergence between the genomes in question. Even in this context, there were still many variants that arose in the mutant stocks that were not present in the original, isogenized *Rev1-ΔCTD*. These variants could represent a number of things. First, if genomic DNA for the homozygotes from the *Rev1-ΔCTD* contained potential heterozygote DNA contamination, this may have resulted in a high level of discerned variation between the stock and the mutants. This can be overcome by a new, careful selection of homozygous *Rev1-ΔCTD* males, followed by re-sequencing, and a

new attempt at the data analysis. Second, genome evolution between the time of sequencing the stock and the time of sequencing the mutants will lead to unavoidable variation. However, in this study, we also had the advantage of the ability to group or subtract non-complementing or complementing mutant strains, respectively. The bioinformatics pipeline output provided manageable candidate lists in two different contexts within this project. First, I identified a strong gene candidate, *polybromo*, which had unique high impact variants in two non-complementing mutant strains (**Table 3-2, 3-3**). Evidence in humans supports a role for a role for the chromatin remodeler Polybromo-1 in the ubiquitination of PCNA, and this made it a strong candidate for a gene that may be involved in DDT, and even TS. I also applied the pipeline to a single mutant (**Table 3-2, 3-4**). This gave a larger list of candidate genes for consideration, but from the list *lodestar* was identified as a high priority candidate for validation (**Figure 3-3**), as it may represent the *D. melanogaster* ortholog of the human HLTF ubiquitin ligase, and thereby would be implicated in error-free lesion bypass in DDT (Gramates et al., 2017; Unk et al., 2008). These candidates, which represent results from only the first few of the EMS-generated mutants from this screen, have the potential to provide significant contributions to our understanding of DDT pathway choice and protein interactions in TS. Moving forward, validating these candidates, and sequencing the other screen-generated mutants, should allow for further contributions to be made. Now that the framework has been streamlined by application in the BLM HELQ project and in this new context, the

identification of candidate variants in all other mutant strains from our EMS-mutagenesis screen should be even more efficient.

In the two projects discussed here, I took a single process and applied it to two unique biological questions. From the bench to the downstream data analysis, much of the framework remained the same, yet it was able to consistently provide valuable, high-resolution information about individual *D. melanogaster* strains. From the WGS data, I was able to both map crossover events and generate mutant lists. In both projects, traditional mapping methods were unnecessary. This was because the subtraction of variant background based on phenotype or complementation data precluded the need for multiple generations of crosses before submitting samples for WGS. In presenting the results of applying these methods to two different studies, I have demonstrated the value of implementing and maintaining a single consistent yet malleable sequencing and data analysis framework for applications in forward genetics research.

REFERENCES

Gramates, L. S., Marygold, S. J., Dos Santos, G., Urbano, J. M., Antonazzo, G., Matthews, B. B., Rey, A. J., Tabone, C. J., Crosby, M. A., Emmert, D. B., Falls, K., Goodman, J. L., Hu, Y., Ponting, L., Schroeder, A. J., Strelets, V. B., Thurmond, J., Zhou, P., Perrimon, N., Gelbart, S. R., Extavour, C., Broll, K., Zytkovicz, M., Brown, N. H., Attrill, H., Costa, M., Fexova, S., Jones, T., Larkin, A., Millburn, G. H., Staudt, N., Kaufman, T., Grumbling, G. B., Cripps, R., Werner-Washburne, M., Baker, P. (2017). FlyBase at 25: Looking to the future. *Nucleic Acids Research*, *45*(D1), D663–D671.

Niimi, A., Hopkins, S. R., Downs, J. A., & Masutani, C. (2015). The BAH domain of BAF180 is required for PCNA ubiquitination. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, *779*, 16–23.

Unk, I., Hajdú, I., Fátyol, K., Hurwitz, J., Yoon, J.-H., Prakash, L., Prakash, S., Haracska, L. (2008). Human HLTF functions as a ubiquitin ligase for proliferating cell nuclear antigen polyubiquitination. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(10), 3768–3773.