2020

Investigating the Role of Centromeric Repeats in Female Meiotic Drive

Jocelyn Crawford

Investigating the Role of Centromeric Repeats in Female Meiotic Drive

A Thesis Presented

by

Jocelyn Crawford

To the Keck Science Department
Of Claremont McKenna, Pitzer, and Scripps Colleges
In partial fulfillment of
The degree of Bachelor of Arts

Senior Thesis in Biochemistry
December 9, 2019

Table of Contents

I. Abstract

Female meiotic drive is an unequal transmission of alleles that arises through the competition of paired chromosomes for inclusion in the egg, resulting in an increase in frequency of the driven alleles regardless of their effect on fitness of the individual. In *Mimulus guttatus* (monkeyflower) second filial generations, driven alleles display transmission advantages resulting in the virtual elimination of recessive homozygotes, while the equivalent lines lacking drive elements conform to traditional Mendelian segregation population ratios. Centromeres have been identified as mechanistic drive elements due to their role in chromosomal segregation during female meiosis, with Mimulus providing the best documented case of centromere-associated female meiotic drive. Here, abundance of centromeric sequence repeats, analogous to centromere size, is quantified and found to be elevated in driver populations, suggesting centromere size as a mechanism for female meiotic drive. A preliminary survey into centromere sequence variation also revealed divergence between driver and non-driver populations, implying centromeric sequence as a secondary mechanistic aspect to drive. The identification of drive related centromere characteristic variation supports a centromere-associated female meiotic drive model, and suggests specific mechanisms for further investigation to elucidate a formidable, but insufficiently understood evolutionary force.

II. Introduction

All eukaryotic organisms rely on inheritance of genetic information and high-fidelity DNA replication for successful cell division and duplication. The centromere, the region responsible for coordinating chromosomal movement during meiosis and mitosis, consists highly repetitive satellite DNA that directs kinetochore formation and attachment to microtubules. The key role centromeres serve in chromosomal segregation, as the bridge between chromosome and spindle during cell division, is reflected by their presence in all eukaryotic chromosomes (Choo 2001; Deininger et al. 2003). An illustration of centromeric importance occurs during female meiosis. During meiosis DNA is replicated, recombined and under goes meiosis one and meiosis two. In females this process results in three polar bodies which degrade over time and one much larger oocyte, containing the genetic material ultimately used for reproduction. Chromosomes are segregated into polar bodies or the oocyte during meiosis II, the final destination determined by centromeres (Zwick et al. 1999).

Some genetic elements take advantage of the competition for inclusion in the oocyte, resulting in unequal transmission and increased frequency of such alleles, regardless of their effect on the fitness of the individual. This non-Mendelian inheritance is female meiotic drive, and has been observed across fungi, insects, plants, and even mammals (Burt and Trivers 2006; Presgraves 2008; Meiklejohn and Tao 2010). Meiotic drive elements are the specific genetic components which depart from Mendelian segregation to increase their own transmission at the expense of homologous loci (Kazazian 2004). These elements have extensive consequences such as reshaping genome structure and speciation process, and producing variation in reproductive fitness, however, the genetic and molecular mechanisms of drive are complex, varied, and, for the most part, remain unclear (Zimmering 1970; Frank 1991).

The potential to influence chromosomal segregation and evolve to increase their presence in the oocyte during female meiosis establishes centromeres as an ideal drive element locus (Chmatal et al. 2014). In a proposed model of centromere drive, centromere expansion was hypothesized to result in an evolutionary advantage, an increased association of microtubules and therefore easier entrance to the oocyte during female meiosis (Figure 1) (Henikoff et al. 2001). In this model centromere expansion is the characteristic most important to drive, which is congruent with cytogenetic observations of larger centromeres in driver populations (Fishman and Saunders, 2008). However, size is not the only centromere characteristic with potential for drive related variation. Centromere binding has been suggested to be sequence dependent, implying sequence deviation as a replacement or additional factor to size in the centromere drive model (Shelby et al. 2000). Determining which centromere characteristic, if either, is the key to the centromere drive model will more broadly inform the deviations from evolutionary biology cannon that have not yet been elucidated. Understanding centromeres and their evolution via female meiotic drive has implications across species, including human cancer therapies (Zhao 2016, Shimo et al. 2008).

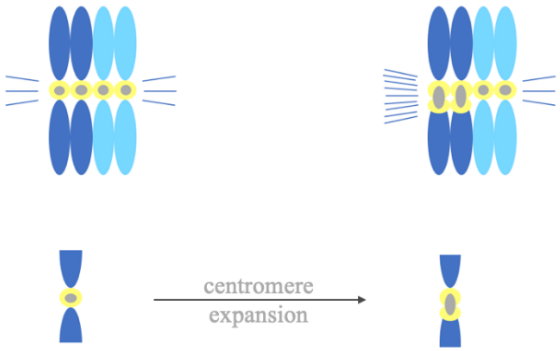


Figure 1. Centromere drive model, adapted from Henikoff et al. (2001). Centromere expansion leads to increased microtubule attachment sites and recruitment, resulting in meiotic advantage during competition for inclusion in oocyte during female meiosis.

The yellow monkeyflower, *Mimulus guttatus* is one of the best documented cases of centromeric drive (Sweigart and Willis, 2003; Fishman and Willis, 2005). This frequently studied organism exhibits within population variation, making it an ideal model system to study drive. Homozygous second filial crosses of *Mimulus guttatus* have displayed non-Mendelian segregation in individuals identified as drivers, while *Mimulus guttatus* non-drivers from the same population did display Mendelian segregation (Fishman and Saunders, 2008). In order to rule out post-meiotic events as the cause for Mendelian equilibrium violation, genome mapping of *M. guttatus* was performed which indicated a variation between driver and non-driver centromeres located on chromosome 11 (Fishman and Willis, 2005). Further cytogenetic imaging conducted by Fishman and Saunders (2008) revealed a qualitatively larger chromosome 11 centromere in drivers compared to non-drivers. This finding is congruent with the suggested model of expansion as the drive associated centromere characteristic (Figure 1).

It is the aim of this study to determine if size, sequence, or both characteristics of centromeres are responsible for determining drive status of an individual, and in doing so develop a robust assay for centromere assessment that can be applied to numerous samples within this model system as well as to other populations potentially demonstrating the same form of centromere associated female meiotic drive. Based on previous findings of centromere characteristics, it is expected that with a sufficiently rigorous assay, both size and sequence will be identified as female meiotic drive mechanistic elements. In order to examine qualities of centromeres, centromere size, or number of tandem repeats, and specific sequence were analyzed in drivers and non-drivers. Molecular qPCR was used to compare centromere size between drivers and non-drivers. Sequences were also obtained using specific chromosome targeting primers developed for qPCR, and analyzed for variation with bioinformatic comparison. Slot blotting was also used as a qualitative size comparison to validate qPCR. This study

III. Materials and Methods

Subjects

Approximately 20 independent lines of *Mimulus guttatus,* or monkeyflowers, from the well documented Iron Mountain (IM) population in Oregon were used in this study. Each line was founded from separate field-collected plant sample seed sets that were maintained by self-fertilization in green house conditions for 5-13 generations (Sweigart et al. 1999; Kelly, 2003; Puzey et al, 2015). Seeds from each line were planted in individual cells and randomized across flats. Previously identified driver and non-driver lines were planted separately, but intermixed within flats. Planted cells were housed in Percival growth chambers on a 16:8 light/dark cycle at 18-22℃ for 6-8 weeks until flowering. All flats were bottom watered every other day to maintain soil moisture. Once mature, leaf tissue samples were removed from the plant and flash frozen in liquid nitrogen, then stored at -80℃ until DNA extraction was performed.

All DNA extractions were performed using the OPS Diagnostics Synergy 2.0 Plant DNA Extraction Kit and associated DNeasy protocol (OPS Diagnostics, USA). All DNA extractions were confirmed with gel electrophoresis and quantified with Invitrogen Qubit 3 Fluorometer and Double Stranded DNA High Sensitivity Reagents according to the manufacturer's protocol (Invitrogen, USA).

Primer Design

PCR primer pairs were designed using the *Mimulus guttatus* v2.0 genome available through the JGI Phytozome v12.1 online database. Target centromere sequences were identified on chromosome 10, 11, and generally throughout the genome. The findings of Fishman and Saunders (2008) suggested increased centromere size on chromosome 11 in drivers as compared to non-drivers. Based on these findings, chromosome 11 represents a driving centromere throughout this study, the general centromere sequence represents an indirect driving centromere quantification, and chromosome 10 represents a control centromere. If centromere size is responsible for determining drive status, when comparing lines of drivers and non-drivers, relative abundance of chromosome 11 specific centromeric sequences, general centromeric sequences, and chromosome 10 specific centromeric sequences are predicted to be drastically more abundant, moderately more abundant, and similar, respectively. Primers were designed using compliment and reverse compliments of half-length target sequences, cut down for optimal PCR conditions, then referenced back to the genome using BLAST to confirm selection of appropriate targets (Appendix B). Replicative PCR was performed using each set of targeted primer pairs and extracted DNA, then gel electrophoresis used to visually confirm amplification of desired target sequence via size verification. Specific PCR conditions available in Table 1. The repetitive nature of the target centromeric repeats required shortened extension time in order to achieve single repeat amplification.

Table 1. Reagents for standard PCR reaction, performed with 58°C annealing temperature and 2 second extension for 35 cycles.

Reagent	Volume Required (μL) for 10 μL Total Sample Volume
ddH_2O	3.35
5x buffer	2
$MgCl_2$ (25 mM)	0.8
dNTPs (2.5 mM)	0.8
BSA (10 mg/mL)	0.5
Forward Primer	0.2
Reverse Primer	0.2
Taq	0.15
Template DNA	2

Centromeric Sequence Comparison

Using template DNA from 3 driver individuals and 3 non-driver individuals, PCR was performed using each set of primer pairs (chromosome 10 specific, chromosome 11 specific, general). Amplification was visually confirmed for each sample with gel electrophoresis, then PCR products were treated with Exonuclease I (Exo I) and Shrimp Alkaline Phosphatase (rSAP) according to protocol presented in Figure 1 to remove extraneous nucleotides or single stranded remnants. These cleaned PCR products were sent for Sanger sequencing via Eurofins Genomics. Sequences were then processed using a script, sequence_processing.sh in Appendix A. This tool was developed to create consensus sequences from paired end reads, eliminating poor quality reads in doing so, then perform a multiple sequence alignment using consensus sequences, which could be used to create a neighbor joining phylogenetic analysis of centromeric sequences.

Table 2. Treatment protocol for PCR products prior to sequencing for 25 μL samples using Exonuclease I (Exo I) and Shrimp Alkaline Phosphatase (rSAP).

Reagent	Volume Required (μL) for 25 μL sample
Exo I	0.8
rSAP	1.6

Temperature (°C)	Time (minutes)
37	15
80	15
8	5

Centromere Sequence Abundance Analysis via qPCR

Once successful replication using targeted primer pairs was confirmed, efficiencies of primers were established for qPCR analysis. All qPCR was performed using reagents and thermal cycling presented in Figure 2. Relevant qPCR primers included the previously designed chromosome 11 specific and general primer pairs, as well as the control housekeeping UBQ5 gene primer pair (Appendix B). The chromosome 10 specific primer pair did not amplify template DNA well under qPCR conditions, resulting in unsuitable amplification curves and exclusion from qPCR analysis. Both the chromosome 11 and general primer pairs were used for qPCR analysis to provide a reference of centromeric repeat abundance which was hypothesized to be related to drive, specifically the chromosome 11 centromeric repeat observed to be larger in drivers by Fishman and Saunders in 2008. Amplification of both the chromosome 11 and general target sequences were thought to be elevated in drivers, however amplification of the chromosome 11 target sequence was expected to be more drastically elevated than the general target between drivers and non-drivers.

Table 3. Protocol for qPCR with 10 µL total well volume and ThermoScientific PowerUp SYBR Green Mastermix.

Reagent	Volume Required (µL) for 10 µL Well Volume
Mastermix	5
Template DNA	4
Forward Primer	0.5
Reverse Primer	0.5

	Temperature (°C)	Time (minutes)
	50	2
	95	3
40x	95	0:10
	58	0:20

DNA sample concentration was standardized to 1ng/µL before dilution for primer efficiency establishment as well as test plates. Test plates were run with 1:100 diluted standardized DNA. Successful qPCR was performed on 3 driver and 3 non-driver DNA templates in quadruplet per primer pair set for the chromosome 11 target sequence, the general centromere target sequence and the UBQ5 control gene primer pairs. The threshold cycle (C_T) refers to the PCR cycle at which the fluorescence signal reaches an arbitrary threshold level. Threshold was set at a point at which the PCR reached exponential amplification. The mean C_T values of each replicate were used for further analysis. In order to normalize the amplification results to the UBQ5 control, difference in threshold cycle (ΔC_T) between UBQ5 target amplification and chromosome 11 specific or general target amplification was calculated. Efficiency corrected fold change, which is directly related to initial target sequence abundance, was calculated using ΔC_T values as dictated by Schmittgen and Livak (2008). Efficiency corrected fold change values were analyzed for significant difference based on drive status using a t-test for independent samples. In order to calculate ΔC_T values, it was necessary that control and test primer sets be run simultaneously on one qPCR plate. Compounded by the need for replicates, the potential for number of samples

compared was limited to 3 drivers and 3 non-drivers. This in turn resulted in small sample size for use when performing t-tests, increasing the possibility of incorrectly finding no significance with t-testing, so a confidence interval of 90% was used, with significant P value set as $p < 0.1$ (Thiese et al, 2016).

Blotting Methods

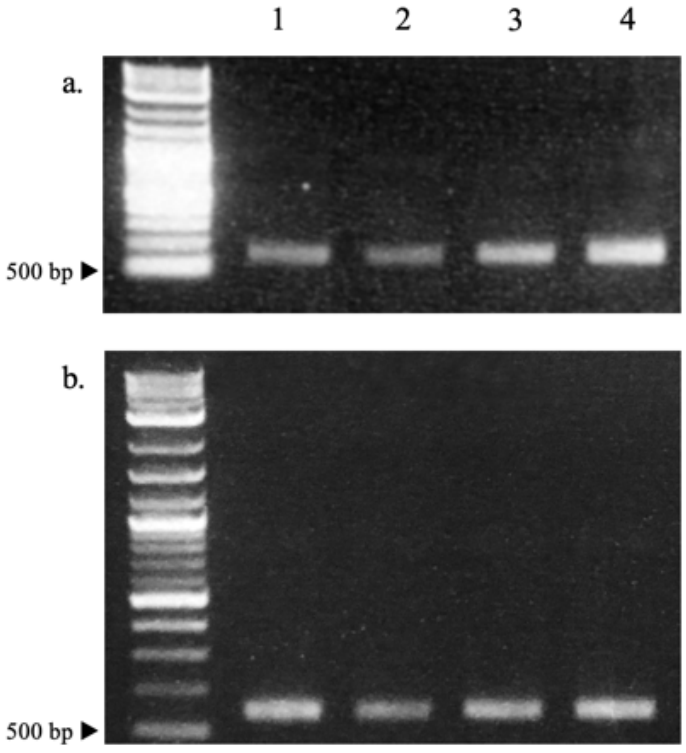
In order to validate sensitive qPCR results, a slot blotting technique that successfully resolved repetitive satellite DNA was adapted for this study to provide a relative centromeric repeat abundance comparison (Khost et al 2017).

Standardized and appropriately diluted DNA samples were denatured at 96℃ for 10 minutes, then quick cooled before hybridization to a positively charged nylon membrane using a vacuum pump equipped Amersham Biosciences PR 648 Slot Blot Filtration Manifold and Ambion NothernMax Prehybridizaton/Hybdridization Buffer according to manufacturer protocols (Amersham Biosciences, UK; Ambion, USA). Ultramer DNA Oligo (4 nmole) biotinylated probes corresponding to chromosome 11 and chromosome 10 specific centromeric sequence targets, as well as a probe previously successful in FISH cytogenetic imaging with sequence as specified by the Fishman and Saunders (2008) study were designed and ordered from Integrated DNA Technologies (Integrated DNA Technologies, USA). Probes were hybridized to the membrane according to the Ambion NorthernMax Prehybridization/Hybridization Buffer manufacturer protocol (Ambion, USA). Probe detection was performed using the Thermo Scientific Biotin Chromogenic Detection Kit according to manufacturer protocol (Thermo Fisher Scientific, USA).

IV. Results

Variation of Targeted Centromeric Sequences

In order to determine centromere sequence variation in terms of potentially contributing to drive elements, the specific sequences of interested must be isolated, analyzed, and compared in reference to predetermined drive status of the source individual. If centromere sequence does influence drive status, drivers will display centromere sequences that are similar to one another and distinct from non-driver centromere sequences. To isolate the genetic regions of interest, purified PCR products of centromeric sequences produced using template DNA from driver and non-driver lines with primer pairs specific to chromosome 10 and chromosome 11, as well as a non-chromosome specific centromere sequence target primer pair, were sent for Sanger sequencing (Figure 2). Due to the repetitive nature of the sequences of interest, only some samples yielded sequencing data suitable for further comparison. Multiple sequence alignment (MSA) of the remaining robust sequences was used to yield neighbor joining trees (Figure 3 and Figure 4).



Lane/Template DNA Individual
1. IM922.10
2. IM138.14
3. IM274.8
4. IM767.12

Figure 2. Agarose gel electrophoresis image of chromosome 10 (a) and chromosome 11 (b) target specific primer pair PCR products. Both primer sets were used with template DNA from multiple driver and non-driver lines. These PCR products are representative of those which yielded sequencing data of high enough quality for MSA analysis and neighbor joining tree comparison, however not all sequenced products are shown here.

In neighbor joining trees, the proximity of branches corresponds to the relatedness of the samples being assessed. Within chromosome 10 specific centromeric sequences, branch clustering was displayed across those sourced from driver identified lines, indicating driver centromeric sequences to be more similar to one another versus the centromeric sequence from a non-driver line (Figure 3). The same clustering pattern was displayed in the tree comparing chromosome 11 specific centromeric sequences (Figure 4). Clustering of drivers and non-drivers based on centromeric sequence is consistent with centromeric sequence influencing drive behavior. Clustering also reflects drivers propagating and sweeping through

a population. This is consistent with the hypothesis that centromere characteristic variation is associated with female meiotic drive displayed in these lines, and suggests sequence as a candidate characteristic related to identifying centromeric drive element potential.

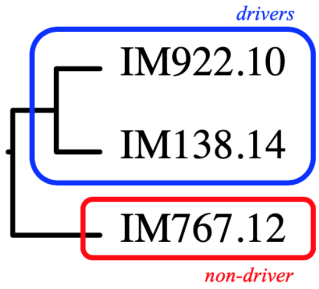
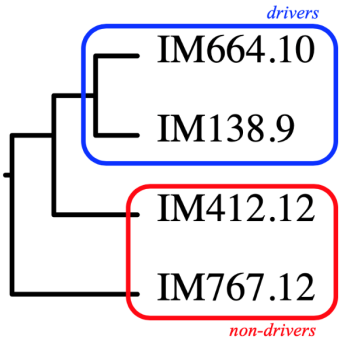


Figure 3. Neighbor-joining tree constructed from MUSCLE multiple sequence alignment using Illumina sequenced PCR products from chromosome 10 specific centromeric targets Identified driver lines indicated in blue, non-driver line indicated in red.

Figure 4. Neighbor-joining tree constructed from MUSCLE multiple sequence alignment using Illumina sequenced PCR products from chromosome 11 specific centromeric targets. Identified driver lines indicated in blue, identified non-driver lines indicated in red.



Quantification of Centromere Size via qPCR Assay

Elucidating centromere size variation and its effect on drive requires the determination of centromeric repeat abundance, in this case synonymous with centromere size, and comparison of centromere size between driver and non-driver lines. If increased centromere size does promote drive as it is thought to, lines exhibiting drive will also exhibit larger centromeres than those of non-driver lines. The chromosome 11 specific centromere is also thought to be a primary driving centromere, and therefore is expected to display more drastic variation in abundance between drivers and non-drivers as opposed to the variation of a general centromere sequence between drivers and non-drivers. Quantification of centromeric repeat abundance was achieved using a qPCR assay.

Raw qPCR amplification curves showed standard amplification patterns, indicating analysis performed using amplification rates is representative of biological occurrences, not faulty replication. Visual confirmation of appropriate amplification curves was required to be confident in further analysis. Analysis of qPCR amplification results confirmed substantial variation in abundance of chromosome 11 specific versus general centromeric repeats, with chromosome 11 specific repeats being more abundant overall (Figure 5a). High variability was also displayed between individual lines, however low error rates reflect low variability within individuals, increasing confidence in the comparisons between samples (Figure 5b,c).

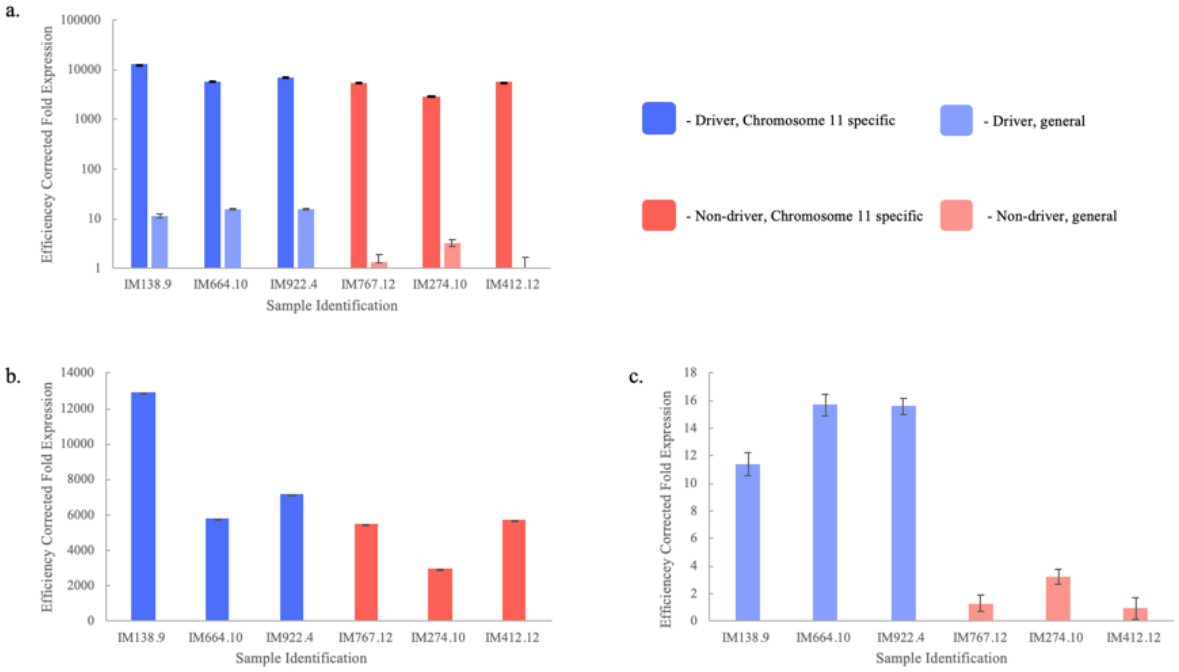


Figure 5. Efficiency corrected fold expression extrapolated from qPCR amplification rates, (for all categories of template DNA type and primer pair specificity n = 4). **a.** Efficiency corrected fold expression of both chromosome 11 specific target sequences and general centromeric sequences shown on a logarithmic scale. **b.** Efficiency corrected fold expression of chromosome 11 specific target sequences using driver and non-driver line template DNA (IM138.9 error ± 0.8517, IM664.10 error ± 0.8445, IM922.4 error ± 0.9460, IM767.12 error ± 0.7528, IM274.10 error ± 0.9175, IM412.12 error ± 0.8972). **c.** Efficiency corrected fold

expression of general centromeric target sequences using driver and non-driver line template DNA (IM138.9 error ± 0.7942, IM664.10 error ± 0.7475, IM922.4 error ± 0.5814, IM767.12 error ± 0.5895, IM274.10 error ± 0.5272, IM412.12 error ± 0.7802).

The abundance of centromeric repeats was significantly different between driver and non-driver lines for both chromosome 11 specific repeats (T-test; $p = 0.0851205$, $df = 4$, $F = 6.04$), and general centromeric repeats (T-test $p = 0.0010615$, $df = 4$, $F = 6.257$) (Figure 5). This assay reinforces centromeres status as loci for female meiotic drive, confirming variation in centromeres associated with drive status. Specifically, this assay quantified centromeric repeat abundance, synonymous with centromere size, which can be verified as a characteristic involved in promoting drive.

Blotting

Comparative slot blotting is a promising option for validating sensitive qPCR results in an effort to create a more vigorous centromere size evaluation tool that eventually can be applied to increased quantities of samples and additional populations. Currently slot blotting has yielded no conclusive results. Successful hybridization of template DNA and target specific probes to nylon membrane has been confirmed with ethidium bromide visualization and rapid spot development of probes with concentrated antibody solution (Figure 6). However, complete membrane antibody detection and development of the DIG labeled probes has not yet been successful or yielded any comparative quantification results. A possible explanation is the removal of the probe during lengthy antibody hybridization procedures. Ethidium bromide aided visualization of membranes reveals undetectable quantities of DNA hybridization at lower concentrations, suggesting increased DNA quantities would improve probe binding rates and aid probe development and detection (Figure 6). Identifying the precise stage and cause of probe loss from membrane, correcting this, and applying the blotting protocol to multiple driver and non-driver line DNA samples will produce a method to validate sensitive qPCR assay findings. Slot blotting in tandem with qPCR will result in a robust and consistent assay tool for assessing centromeric repeat abundance within this, and other populations.

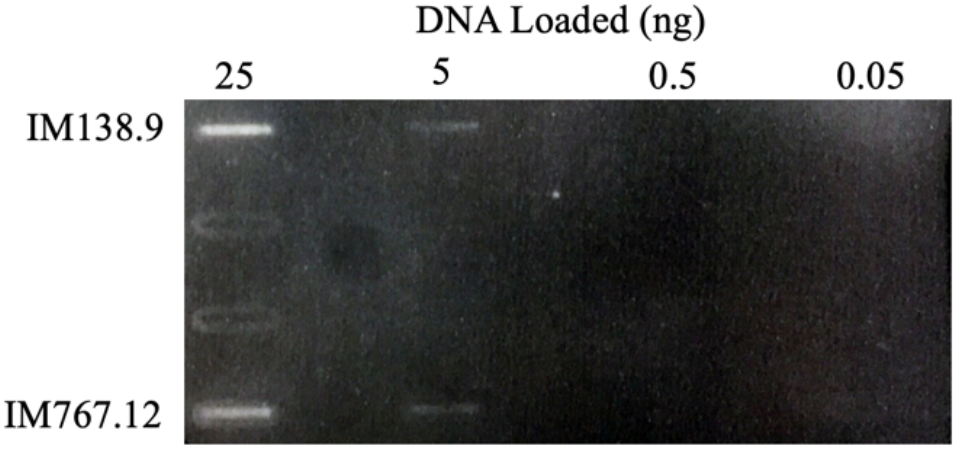


Figure 6. Preliminary slot blot visualized with ethidium bromide displaying hybridization of denatured genomic DNA and probe to nylon membrane across decreasing concentrations prior to probe development, for one driver (IM138.9) and one non-driver (IM767.12).

V. Discussion

In female meiosis, the "stronger" centromere wins the competition for advantageous orientation during meiosis, and in turn inclusion in the oocyte. There is fairly substantial evidence supporting a model in which increased size is the relevant centromere characteristic for establishing strength in monkeyflowers. Cytogenetic imaging of driver and non-driver lines revealed larger centromeres on chromosome 11 in drivers. Further support of this finding comes from genome mapping indicating variation between driver and non-driver centromeres on chromosome 11, as well as ruling out post-meiotic events as cause of distorted transmission ratios. Centromeric expansion is also favored as the relevant factor in models of centromere and histone co-evolution (Henikoff, 2001; Dawe and Henikoff, 2006; Malik and Bayes, 2006). Sex determination has also been connected to centromere size and female meiotic drive in birds (Rutkowska and Badyaev, 2007). However, quantification of centromeric repeat abundance in the often studied *Mimulus* system has yet to confirm the qualitative results presented in previous works.

Although many models of drive favor size as the major decider of centromeric "strength", there is evidence to suggest sequence plays an important role as well. Centromere binding has been established as sequence dependent, contributing to its coevolution with histones and development of "selfish" transmission characteristics (Shelby et al. 1997; Keith et al. 2000). Centromeric repeats evolve exceptionally rapidly, even across closely related species, most likely due to competitive coevolution with histones, further supporting involvement of sequences as relevant in centromeric drive models (Haaf and Willard, 1997; Csink and Henikoff, 1998). Due to the repetitive nature and incredible diversity of centromeric repeats, extensive sequence comparison in correlation with drive status has not yet been performed with conclusive results.

Here, centromere size and sequence were analyzed for variation between drivers and non-drivers. Similar to the previous works in this field, obtaining data sets with large sample sizes was challenging because of the repetitive nature of this satellite DNA region. Comparison of few representative individuals showed sequence variation related to centromere source line drive status. This finding is consistent with the claim that centromere sequence defines centromeres as meiotic drive elements. This study was successful in quantifying centromeric repeat abundance in driver and non-driver samples, confirming quantitative cytogenetic imaging results suggesting larger centromeres in driver populations, as was expected (Fishman and Saunders 2008). While quantitative appraisal of centromeres was obtained, the qPCR assay used was sensitive and implied the necessity of an additional validation technique in order to develop a tool capable of comparing centromeres of numerous individual samples. This was addressed with a slot blot procedure.

The blotting validation methodology is still in early stages, however based on success in previous studies and some preliminary results showing compatibility with the repetitive centromeric repeats of interest, is promising as a tool to contribute to the centromere assessment protocol developed here (Khost et al. 2017). The potential of this study was limited by small sample sizes, in order to fulfill the goal of rigorous comparison of centromeres across *Mimulus guttatus* drivers and non-drivers, application of the methods described here to more individuals is required. While these methods are promising for use as a general centromere associated drive evaluation tool, further optimization and sample application are necessary. With the completion of these further steps, centromeres can be

understood better than previously possible in the context of the complex and evolutionarily significant context of female meiotic drive.

References

Chmátal, L., Gabriel, S.I., Mitsainas, G.P., Martínez-Vargas, J., Ventura, J., Searle, J.B., Schultz, R.M. and Lampson, M.A., 2014. Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. *Current Biology*, *24*(19), pp.2295-2300.

Choo, K.A., 2001. Domain organization at the centromere and neocentromere. *Developmental cell*, *1*(2), pp.165-177.

Csink, A.K. and Henikoff, S., 1998. Something from nothing: the evolution and utility of satellite repeats. *Trends in Genetics*, *14*(5), pp.200-204.

Dawe, R.K. and Henikoff, S., 2006. Centromeres put epigenetics in the driver's seat. *Trends in biochemical sciences*, *31*(12), pp.662-669.

Deininger, P.L., Moran, J.V., Batzer, M.A. and Kazazian Jr, H.H., 2003. Mobile elements and mammalian genome evolution. *Current opinion in genetics & development*, *13*(6), pp.651-658.

Fishman, L. and Willis, J.H., 2005. A novel meiotic drive locus almost completely distorts segregation in Mimulus (monkeyflower) hybrids. *Genetics*, *169*(1), pp.347-353.

Fishman, L. and Saunders, A., 2008. Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science*, *322*(5907), pp.1559-1562.

Frank, S.A., 1991. Divergence of meiotic drive-suppression systems as an explanation for sex-biased hybrid sterility and inviability. *Evolution*, *45*(2), pp.262-267.

Haaf, T. and Willard, H.F., 1997. Chromosome-specific α-satellite DNA from the centromere of chimpanzee chromosome 4. *Chromosoma*, *106*(4), pp.226-232.

Henikoff, S., Ahmad, K. and Malik, H.S., 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, *293*(5532), pp.1098-1102.

Kazazian, H.H., 2004. Mobile elements: drivers of genome evolution. *science*, *303*(5664), pp.1626-1632.

Keith, K.C. and Fitzgerald-Hayes, M., 2000. CSE4 genetically interacts with the Saccharomyces cerevisiae centromere DNA elements CDE I and CDE II but not CDE III: implications for the path of the centromere DNA around a Cse4p variant nucleosome. *Genetics*, *156*(3), pp.973-981.

Kelly, J.K., 2003. Deleterious mutations and the genetic variance of male fitness components in Mimulus guttatus. *Genetics*, *164*(3), pp.1071-1085.

Khost, D.E., Eickbush, D.G. and Larracuente, A.M., 2017. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster. *Genome research*, *27*(5), pp.709-721.

Malik, H.S. and Henikoff, S., 2001. Adaptive evolution of Cid, a centromere-specific histone in Drosophila. *Genetics*, *157*(3), pp.1293-1298.

Malik, H.S. and Bayes, J.J., 2006. Genetic conflicts during meiosis and the evolutionary origins of centromere complexity.

Meiklejohn, C.D. and Tao, Y., 2010. Genetic conflict and sex chromosome evolution. *Trends in Ecology & Evolution*, *25*(4), pp.215-223.

Presgraves, D.C., 2008. Sex chromosomes and speciation in Drosophila. *Trends in Genetics*, *24*(7), pp.336-343.

Puzey, J.R., Willis, J.H. and Kelly, J.K., 2015. Whole genome sequencing of 56 Mimulus individuals illustrates population structure and local selection. *bioRxiv*, p.031575.

Rutkowska, J. and Badyaev, A.V., 2007. Meiotic drive and sex determination: molecular and cytological mechanisms of sex ratio adjustment in birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1497), pp.1675-1686.

Shelby, R.D., Monier, K. and Sullivan, K.F., 2000. Chromatin assembly at kinetochores is uncoupled from DNA replication. *J Cell Biol*, *151*(5), pp.1113-1118.

Shimo, A., Tanikawa, C., Nishidate, T., Lin, M.L., Matsuda, K., Park, J.H., Ueki, T., Ohta, T., Hirata, K., Fukuda, M. and Nakamura, Y., 2008. Involvement of kinesin family member 2C/mitotic centromere-associated kinesin overexpression in mammary carcinogenesis. *Cancer science*, *99*(1), pp.62-70.

Sweigart, A., Karoly, K., Jones, A. and Willis, J.H., 1999. The distribution of individual inbreeding coefficients and pairwise relatedness in a population of Mimulus guttatus. *Heredity*, *83*(5), p.625.

Sweigart, A.L. and Willis, J.H., 2003. Patterns of nucleotide diversity in two species of Mimulus are affected by mating system and asymmetric introgression. *Evolution*, *57*(11), pp.2490-2506.

Trivers, R. and Burt, A., 2006. Genes in conflict: The biology of selfish genetic elements.

Zhao, H., Zhu, X., Wang, K., Gent, J.I., Zhang, W., Dawe, R.K. and Jiang, J., 2016. Gene expression and chromatin modifications associated with maize centromeres. *G3: Genes, Genomes, Genetics*, *6*(1), pp.183-192.

Zimmering, S., Sandler, L. and Nicoletti, B., 1970. Mechanisms of meiotic drive. *Annual review of genetics*, *4*(1), pp.409-436.

Zwick, M.E., Salstrom, J.L. and Langley, C.H., 1999. Genetic variation in rates of nondisjunction: association of two naturally occurring polymorphisms in the chromokinesin nod with increased rates of nondisjunction in Drosophila melanogaster. *Genetics*, *152*(4), pp.1605-1614.

David M. Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, and Daniel S. Rokhsar, **Phytozome: a comparative platform for green plant genomics**, *Nucleic Acids Res.* 2012 40 (D1): D1178-D1186

Appendix A

sequence_processing.sh

```
# Jul 12 2019
# Working script to process short Sangar sequencing reads from PCR
products, F and R but not in fasta or fastq format yet

# Copy files from computer onto the server
# Put all sequencing result files into a single directory, then copy
directory onto local directory on server
# Do this before starting script in separate window using sftp

###############SFTP COMMANDS##############
## sftp jcrawford20@cbsulogin2.tc.cornell.edu
## enter password
## now should be in jcrawford20 home directory, can confirm with pwd
## put -r
Users/jocelyncrawford/Documents/Findley_Research/Sequences/Seq"batch_nu
m"_processing
## files uploaded should be in .phd.1 format but should not need any
further processing
## when calling script put batch number in command line

# start script here
set -e ## terminates script if command fails
set -u ## aborts script if variable's value not set
set -o pipefail ## terminates script if pipe fails

for sample in "$@"; do
#################MOVE SEQUENCES INTO WORKING DIRECTORY###########
# upload all phd.1 files in one directory according to batch number
# move into that directory
cd /workdir/genomics2018/jcrawford20/sequences/Seq"$sample"_processing

# later for drawgram program, will need font file in all directories so
copy that in now
cp /workdir/genomics2018/jcrawford20/sequences/fontfile
/workdir/genomics2018/jcrawford20/sequences/Seq"$sample"_processing

#################PUT FILES IN FASTA FORMAT##################
# need fasta of sequences and quality for F and R

# delete headers (need the -i to make changes to file in place)
sed -i '1,19d' *.phd.1

# delete footer lines "END SEQUENCE / END DATA"
sed -i '/^E/d' *.phd.1

# replace all spaces with tab
sed -i 's/ /\t/g' *.phd.1

#create a new file for sequences, column 1
```

```
find -type f -iname "*.phd.1" -exec awk '{print $1 >(FILENAME "-
sequences")}' {} \;

# create new file for quality, column 2
find -type f -iname "*.phd.1" -exec awk '{print $2 >(FILENAME "-
quality")}' {} \;

# replace all enter with nothing for sequence files
sed -i ':a;N;$!ba;s/\n//g' *-sequences

# replace all enter with spaces for quality files
sed -i ':a;N;$!ba;s/\n/ /g' *-quality

# insert header in fasta format, starting with ">", also quality and
sequence headers must match
# this works to add header with text following ">" as file name minus
whatever suffix is specified after "-s"
      #header for sequence files
for f in *.phd.1-sequences; do
    base=$(basename -s  .phd.1-sequences "$f")
    sed -i "1 i\>$base" "$f"
done

      #header for quality files
for f in *.phd.1-quality; do
    base=$(basename -s  .phd.1-quality "$f")
    sed -i "1 i\>$base" "$f"
done

#################FILES ARE NOW IN FASTA FORMAT############
#merge fastas into one large sequence file and one large quality file
#for sequences forward
cat *"primercode"F*.phd.1-sequences > forward_sequences.fa

#for quality forward
cat *"primercode"F*.phd.1-quality > forward_quality.fa

#for sequences reverse
cat *"primercode"R*.phd.1-sequences > reverse_sequences.fa

#for quality reverse
cat *"primercode"R*.phd.1-quality > reverse_quality.fa


#################TURN FASTA INTO FASTQ#####################

# merge sequence and quality files using QUIIME script
# set environment
export PATH=/programs/miniconda2/bin:$PATH
source activate qiime1

# command to merge into fastq, headings must match for merging
# "-f" specifies input sequence file
# "-q" specifies input quality file
```

```
# for forward sequences
convert_fastaqual_fastq.py -f forward_sequences.fa -q
forward_quality.fa

# for reverse sequences
convert_fastaqual_fastq.py -f reverse_sequences.fa -q
reverse_quality.fa

#outputs will be forward_sequences.fastq and reverse_sequences.fastq
#After you are done
conda deactivate

###############USE PEAR TO MERGE PAIRED END READS#############
# add latest version of pear to path, can use program just by typing
name in prompt
export PATH=/programs/pear:$PATH

# command to merge paired ends, produces consensus
pear -f forward_sequences.fastq -r reverse_sequences.fastq -o consensus

#####CHECK THIS this might just assemble to forward reads, compare
against manually done version

# all of the aligning programs require FASTA format, our files are in
FASTQ so using QUIIME script to change format
########################TURN FASTQ INTO
FASTA##############################
# separate sequence and quality files using QUIIME script
# set environment
export PATH=/programs/miniconda2/bin:$PATH
source activate qiime1

# converting consensus FASTQ to FASTA only
# -f is input file
# -c specifies fastq to fasta + quality files

# for consensus
convert_fastaqual_fastq.py -f consensus.assembled.fastq -c
fastq_to_fastaqual

# outputs will be consensus.assembled.fna
# After you are done
conda deactivate

########################COMPARE SEQUENCES#############################
# want to compare consensus sequences to one another now using MUSCLE
to create N-J tree
# if muscle isn't already added to path use command:
export PATH=/programs/muscle3.8.31:$PATH

# make alignment
muscle -in consensus.assembled.fna -out consensus.afa
# then to make tree
```

```
muscle -maketree -in consensus.afa -out consensus.phy -cluster
neighborjoining

# output is in Newick format

############################VISUALIZE
TREE################################
# using drawgram from PHYLIP to visualize tree from Newick format file

# add drawgram to path
export PATH=/programs/phylip-3.696/exe:$PATH

# use drawgram program
drawgram
# specifiy tree file input name, output from muscle

consensus.phy

# will need a font file within the current directory named "fontfile"
or specify name of font file
# font files downloaded from PHYLIP, preference saved as fontfile

# defult output file name will be "plotfile"
# rename plotfile
mv plotfile Seq"$sample"_tree

# tree is now complete and drawgram output can be visualized as a PDF
IF moved to local device that can view PDF format
# use sftp to "get" tree files and visualize/compare

echo DONE with "$sample"
done
```

Appendix B
Primers and Associated Target Sequences

```
>chr11seq
acaatagataaaatagatacaatggagtgtagcaataatactcgacgtttacgtttacgtggatttttgtt
tggattaatttatatttctgtttttattggtggatttggtatcttttaggtacaaatcgaaaattaggtgt
ccaggaagcaaaaatgatcaaagggtaagaaaagagttgcaattaggaaaataaaatcaaattagaagaga
gaatccctcaacatctttggtagtgtccacgaagtcgagcataactctctcatccggactccaaatcggag
tggttccggtggcattagaaagctatttcagtgggctacaattcccttctaacgtcaaaattccaaattcg
gactcgaacatggtcataattggataacaaa
>primers_chr11seq
>forward
tcaaagggtaagaaaagagttgca
>reverse
ctttctaatgccaccggaacc

>chr10seq
tttacgttgatttttttgtggattaatttatatttctatttgaattggtggatttggtatcttataggtac
aaatcaaaaaataggtgtcaaggaagcaaaaatgatcaaagagtaagaaaagtgttggaactaggacaata
aaatcaaattagaagagagaatccctcaacatctttggtagtgtacgtccacgaagtcgagcataactttc
tcaaccggactccaagtcaactggttctggcggcattagaaagctatttcagctggctacaattcctgtct
aacggcaaaaatccaaattaggactcgaacatggtcaaaattg
>primers_chr10seq
>>left
gagtaagaaaagtgttggaactagg
>>right
agctttctaatgccgccaga

>generalseq
aaagatacaatagagtgtagcaataatactcgacgtttacgtttacgtggatttttgcttggattaattta
tatttctatttttattggtggatttggtatctttttggtacaaatcgaaaataaggtgtccaggaagcaaa
aatgatcaaagggtaagaaaagagttgcaattaagaaaataaaatcaaattagaagagagaatccctcaac
atcttaggtattgtccacgaagtcgagcataactctctcatccggactccaaatcgactggttccgatagc
attagaaagctatttcaggggggctacaattcccgtctaacgtcaaaattccaaattcgaactcgaacatgg
tcaaaattggataacaaa
>primers_genseq
>>left
acgtttacgtttacgtggattttgg
>>right
Tggagtccggatgagagagt

>FISH_probe from Fishman and Saunders 2008 cytogenetic imaging study
>>left CTTCCTGGACACCTAATTTTCG
>>right CAAAGGGTAAGAAAAGAGTTGC
```