TESTING THE ROLE OF *D. MELANOGASTER MATERNAL HAPLOID* IN A *D. SIMULANS* X *D. MELANOGASTER* HYBRID CROSS

Honors Thesis
Presented to the College of Agriculture and Life Sciences,
Cornell University
in Partial Fulfillment of the Requirements for the
Biological Sciences Honors Program

by Sahana Natesan
November 2019

Advisors:
Daniel Barbash, Department of Molecular Biology and Genetics, Cornell University
Dean M Castillo, Department of Biology, University of Utah

Table of Contents

Abstract

Hybrid incompatibility (HI) (such as hybrid sterility or lethality) is a reproductive isolation barrier that contributes to speciation. Genes have been identified whose interaction causes HI; however, identifying the factors (i.e. maternal genes, small RNAs, etc.). HI is essential to understanding how these genes function to affect hybrid development. Previous studies have shown that when *D. melanogaster* female parents are mated with *D. simulans* male parents, the interaction of two genes - *Hybrid male rescue* and *Lethal hybrid rescue* - causes hybrid male lethality. When a *D. simulans* female parent is crossed with a *D. melanogaster* male parent we observe the opposite outcome: hybrid female offspring die in the embryo stage while hybrid males live. At this stage, embryonic cells fail to undergo mitosis appropriately. During anaphase, the X chromatids segregate partially or not at all. This abnormal segregation is attributed to the 359-bp satellite DNA in *D. melanogaster* which maps to the *Zygote hybrid rescue* (*Zhr*) locus. Since we know that in a pure *D. melanogaster* cross, all of the offspring live, we hypothesize that there is a factor which regulates *Zhr* to allow for normal mitosis to occur. Maternal haploid (*Mh*) is maternal factor which is an important protease involved in decondensation of the paternal genome during zygote formation. To test whether maternal haploid is a potential factor which drives HI, we created transgenic strains of *D. simulans* which contained the *D. melanogaster* maternal haploid protease. I then mated female transgenic flies with *D. melanogaster* males and compared the hybrid progeny results to that of the control cross. I found that with the presence of *D. melanogaster* maternal haploid in the genome of the *D. simulans* fly, the female hybrid viability rate was higher than among the female hybrid progeny of normal *D. simulans* flies. This shows that maternal haploid is a factor that contributes to hybrid viability.

Introduction

Evolutionary history shows that over time populations evolve through a process known as speciation. Species are often reproductively isolated by barriers such as time or environment which discourage hybridization - the process by which two organisms of different species mate with one another to produce offspring. The two parental genomes of these hybrids may be incompatible leading to hybrid sterility or lethality. Hybrid incompatibility reduces the exchange of genetic material between species and thus is a reproductive isolating barrier to speciation. Past studies have shown that the reason hybrids experience these incompatibilities is due to alleles at different genetic loci that tend to not function well together, or in other words, negatively interact (Johnson, 2000). Specific genes which are known to play a role in hybrid incompatibility have been identified and sequenced - many of these genes are found in the *Drosophila* genus. Studying the interaction of these genes allows us to better understand the nature of hybrid sterility and lethality. Theodosius Dobzhansky and Hermann Joseph Muller, collectively known as Dobzhansky-Muller, performed genetic analyses on Drosophila hybrids to study the evolution of hybrid incompatibility (HI). They concluded that an interaction between two alleles at different loci in the parent genomes results in negative effects to the progeny.

As an example of the Dobzhansky-Muller model, two genes - *Hybrid male rescue* (*Hmr*) and *Lethal hybrid rescue* (*Lhr*) - cause hybrid male lethality when *Drosophila melanogaster* and *Drosophila simulans* flies mate to produce progeny. In this case, *D. melanogaster* female parents are mated with *D. simulans* male parents (Brideau, 2006). If we flip the direction of this cross, where a *D. simulans* female parent is crossed with a *D. melanogaster* male parent we observe the opposite outcome: the hybrid female offspring die in the embryo stage while hybrid males do not. At this stage, embryonic cells in females fail to undergo mitosis appropriately. During anaphase, the X chromatids segregate partially or not at all (Ferree and Barbash, 2009).

This abnormal segregation is attributed to a block of satellite DNA in *D. melanogaster*, known as *Zygote hybrid rescue* (*Zhr*). It has a repeating unit of 359 base pairs and is found near the centromere of the X chromosome (Sawamura, 1993). In hybrid female embryos, the *D. simulans* X chromatids segregate normally leading to the conclusion that the lagging chromatin in the hybrid female is caused by irregular segregation of X chromatids in the *D. melanogaster* male parent (Ferree and Barbash, 2009).

Consider the cross between a *D. melanogaster* male parent and a *D. melanogaster* female parent: all of the offspring live. As mentioned, the 359-bp block causes abnormal separation of chromatids by *D. melanogaster* X chromosomes in all *D. melanogaster* flies and thus the offspring of a pure *D. melanogaster* cross should not survive. However, since we know that the offspring are viable, we know that there must be some factor that suppresses the function of *Zhr* to allow for normal mitosis to occur. The fact that hybrid females are lethal in a cross between a *D. simulans* female parent and a *D. melanogaster* male parent yet viable in the reciprocal cross affirms that this factor is a maternal effect. Further, unlike male flies who can only contribute their genome to a zygote during fertilization, female parents are able to give their genomes additional maternal factors such as small RNAs and proteins.

Maternal haploid (*Mh*) is an important protease involved in decondensation of the paternal genome during zygote formation. In *D. melanogaster* it appears to be enriched at the 359-bp region indicating that it may be responsible for suppressing the *Zhr* gene in offspring. In *Mh* mutant flies, *D. melanogaster* offspring experience instability at the 359-bp satellite - sister chromatids at that location do not completely segregate (Tang, 2017). Given that without

Maternal haploid, progeny are prone to abnormalities in paternal genome decondensation or *Zhr* suppression, *Mh* is a good candidate maternal factor whose absence could cause hybrid incompatibility.

Although *Zhr* is not found in *D. simulans*, these flies contain two copies of *Maternal haploid*, but little is known about their function. Among *Drosophila* species, it is quite uncommon for duplicate genes to evolve and typically, between duplicate genes, one copy is expected to evolve neutrally while the other tends to lose functionality (Assis, 2014). The *Mh* found in *D. melanogaster* (hereby referred to as Mel mh) and the *Mh* found in *D. simulans* (hereby referred to as Sim mh_{Dup1} and Sim mh_{Dup2}) are divergent genes: though they may have originated from a common ancestor, over time, these genes have acquired mutations which could make them quite different from one another and likely that Sim mh and Mel mh have different roles. This serves as an additional reason for female hybrid lethality in the cross between a *D. simulans* female parent and a *D. melanogaster* male parent.

We want to test the candidate gene, Mel mh, to determine whether its absence from the *D. simulans* genome contributes to female hybrid lethality in progeny of *D. melanogaster* male parents and *D. simulans* female parents. Based on the assumption that Mel mh and Sim mh are different from one another, we theorize that if the *D. simulans* female parent had Mel mh in her genome, the female hybrid offspring produced with *D. melanogaster* males should survive. Since Mel mh is usually enriched at the 359-bp region of the X chromosome pericentromere, *Zhr* should be suppressed and chromatid segregation should occur normally. In the *D. melanogaster* male x *D. simulans* female cross (referred to as the control cross), there should be little to no female hybrid offspring compared to the *D. melanogaster* male x *D. simulans* + Mel mh female cross (referred to as the experimental cross) where there should be a significant increase in the number of alive female hybrid offspring. *D. simulans* on its own tends to be “leaky” - a term used when describing strains of flies which produce some flies of an unexpected phenotype. In this case, though we expect that in the *D. melanogaster* male x *D. simulans* female cross that no female progeny should develop, it is possible that we may see a few. Once we have counted the number of hybrid progeny from both the control cross and the experimental cross, we can use a Chi Square test to test for statistical significance between the values.

In addition, we want to quantify how divergent Mel mh, Sim mh_{Dup1} and Sim mh_{Dup2} are. Since Sim mh exists as two copies in the *D. simulans* genome, we already know that an array of mutations may have occurred because duplicate genes accrue mutations making them more prone to becoming inactive (Georgia, 2014). However, to appropriately quantify the divergence from Mel mh, we can align the gene sequences of Mel mh, Sim mh_{Dup1} and Sim mh_{Dup2} and count the number of synonymous and nonsynonymous substitutions to obtain dN/dS ratios which we can compare to other genomic data to tell us how the genes are evolving.

Finally, we want to prove that Mel mh is actually being expressed in a *D. simulans* background. We begin by injecting the fly with a plasmid which contains an attP site and the Mel mh. The catalyzation of the injected attP with the inherent attB site allows for integration of the Mel mh into the *D. simulans* genome. To test that Mel mh is being expressed, we can use RT PCR to amplify specific genomic sequences where we expect to see Mel mh. We expect to find *Mh* in the ovaries of the female flies as it is a maternal gene. We can extract the RNA from the ovaries of female flies and design primers which will bind to specific areas in the genome where we expect to see Mel mh in the *D. simulans* genome. Then by performing a restriction digest on the PCR bands where we see expression, we can see if one or both duplicates of Sim mh are expressed in the genome of *D. simulans* flies.

Methods and Materials

I. Nomenclature

D. melanogaster

Abbreviation of *Drosophila melanogaster,* one of two main *Drosophila* species used in these experiments. The sub-strain used is Canton-S.

D. simulans

Shorthand version of the name *Drosophila simulans* and is the second of two main *Drosophila* strains used in these experiments. There are two sub-strains of *D. simulans* used: sim1029 and sim1048.

Mh

Maternal haploid gene

Mel mh

Maternal haploid ortholog from *D. melanogaster*

Sim mh (Sim mh_{Dup1} and Sim mh_{Dup2})

Maternal haploid ortholog(s) from *D. simulans*

D. simulans + Mel mh

Drosophila simulans flies which have *D. melanogaster maternal haploid* integrated into their genome. The two sub-strains of *D. simulans* (sim1029 and sim1048), after the injection with Mel mh are referred to as sim1029mh and sim1048mh.

II. Drosophila Lines

The fly lines *Drosophila melanogaster* and *Drosophila simulans* have long served as model species when studying evolution and more specifically, hybrid incompatibility (Barbash, 2010). In our experiments, we sought to determine whether the *Mh* gene could regulate the 359-

base pair locus found in *Drosophila melanogaster*. In order to test the effects of this gene, we first needed to inject the *D. simulans* flies with the maternal haploid gene found in *D. melanogaster*. We obtained two strains of *D. simulans* flies in which the attP site was integrated and mapped to two different genomic locations – 3R and 2L (Stern, 2017). The *D. simulans* strain with attP at the 3R chromosomal location is referred to as sim1029 and the strain with attP at 2L as sim1048. These specific strains were chosen because it was shown that all flies in the stock were still viable even with the integration of the attP sites. In addition, the integration of attB plasmids into these attP landing sites was relatively efficient (Stern, 2017).

Once we had obtained the 1029 and 1048 *D. simulans* flies, we needed to inject these flies with the *D. melanogaster Maternal haploid* plasmid (performed by an external lab). These *D. simulans* + Mel mh fly strains were known as sim1029mh and sim1048mh. The injected plasmid also contained a w+ marker which allowed us to determine whether the further generation progeny were homozygous, heterozygous or neither. In other words, since the the *D. simulans* stocks were white eyed, if an offspring had white eyes, we concluded that it had not received any of the Mel mh from its *D. simulans* + Mel mh parents. Likewise, if the progeny had light red eyes, we could classify it as heterozygous as it had only one copy of Mel mh. If the progeny had dark red eyes, it was homozygous and has received two copies of Mel mh, one from each parent. From *D. simulans* (sim1029 or sim1048) or *D. simulans* + Mel mh (sim1029mh or sim1048mh), we collected virgin female flies for our experiments.

The male flies in this experiment came from the widely used Canton-S stock, which is a sub-strain of *D. melanogaster*. Not only has this strain been successfully used in the Barbash Lab at Cornell University, several studies have shown that it mates well with *D. simulans*.

III. Expression of the *D. melanogaster* mh Allele in *D. simulans* + Mel mh

We wanted to be sure that Mel mh was in fact being expressed in the *D. simulans* fly in which it had been integrated. A challenge is that *D. simulans* itself has its own form of *Maternal haploid* (referred to as Sim mh). In order to differentiate the multiple *Mh* alleles (either Sim mh or Mel mh) we created a sequence alignment of the Mel mh gene with the *D. simulans* orthologs (found on FlyBase). When we aligned Mel mh and Sim mh, we discovered that there was a duplicate copy of Sim mh in *D. simulans*.

Maternal Haploid in *D. Simulans*

We determined the ratio between synonymous to non-synonymous substitutions to quantify divergence between Mel mh and the Sim mh duplicate orthologs. flyDIVaS is an online database which has the dN/dS ratio calculated for over 10,000 different genes – however, the score for *Maternal haploid* is incorrect because the alignment between Sim mh and Mel mh was done incorrectly. Thus, we designed a method to properly align Mel mh and Sim mh to then calculate a dN/dS ratio.

We began by obtaining the genome sequences of Mel mh, Sim mh_{Dup1} and Sim mh_{Dup2} from FlyBase. Using the Clustal Omega program, we were able to create the alignment and then use MEGA X to remove the intron sequences from the Mel mh genome sequence. These noncoding regions of Mel mh were identified by lining up the coding sequence of Mel mh with the full gene sequence of Mel mh and identifying areas where the nucleotides dd not match each other. Then we fed the alignment into SnapGene Viewer where we identified single nucleotide polymorphisms (SNPs) – areas where the sequences differed from one another. Finally, through the SNAP program we calculated the dN and dS score and ratio for each pair of genes (i.e. Mel mh and Sim mh_{Dup1}, Mel mh and Sim $mh_{Dup2,}$ and Sim mh_{Dup1} and Sim mh_{Dup2}). In order to verify our method, we repeated our methodology for *Lethal hybrid rescue* (*Lhr*), for which the dN/dS value has previously been appropriately calculated on flyDIVaS.

The alignment was also useful to design primers for the RT PCR experiments. First, we wanted to find primers which would bind to the respective *Mh* genes: Mel mh, Sim mh_{Dup1} and Sim $\text{mh}_{\text{Dup2.}}$ Then, in order to differentiate between the two duplicates of Sim mh, we wanted to create restriction sites within those primers which would be polymorphic as to cleave one duplicate and not the other. By aligning the three sequences, we could find areas where one of the Sim mh duplicates matched with the Mel mh sequence and the second duplicate did not.

To find *D. simulans* primers we identified an area in which the sequences of both duplicates of *D. simulans* mh were the same but there were SNPs in the *D. melanogaster* mh. We repeated these experiments to find primers for *D. melanogaster*. After locating primer regions, we used the Primer3 program to design forward and reverse *D. simulans* specific and *D. melanogaster* specific primers which would flank those cut sites. The purpose of creating *D. simulans* specific primers and *D. melanogaster* specific primers was to amplify specific parts of the DNA (through polymerase chain reactions – PCR) to compare bands obtained from gel electrophoresis between flies containing only Mel mh, those containing Sim mh and those containing both Mel mh and Sim mh.

After the primers were designed, we collected RNA from our samples to undergo reverse transcription into cDNA to use for PCR. RNA was collected from the ovaries of female flies from each genotype – sim1029, sim1029mh, sim1048, sim1048mh and Canton-S. The Canton-S RNA was used as a control for *D. melanogaster* mh while the sim1029 and sim1048 RNA were used as controls for *D. simulans* mh in each strain respectively. We chose to collect from ovaries because we know that maternal haploid is a maternal gene which is passed on to offspring. The

Qiagen Kit and corresponding protocol were used to dissect 20 ovaries from each of the 5 genotypes. The RNA content of the ovaries was measured using the Qubit software and with those known concentrations, the appropriate amounts of primer mix and cDNA synthesis mix were added to convert the RNA to cDNA. For each sample of RNA, we created a sample of cDNA which had the reverse transcriptase enzyme (+RT) and one sample of cDNA which did not (-RT). This allowed us to use the -RT sample as a control because if the RNA had never been converted to cDNA, the PCR experiments to come will not be able to amplify any of those samples. Thus, we could account for contamination if we did see bands front the –RT samples.

To our cDNA samples we added the appropriate primers to prepare for PCR. Each sample was run once with *D. melanogaster* primers and once with *D. simulans* primers. We used the GoTaq buffer and enzyme. The PCR conditions were: DNA degradation – 95 degrees C for 2 minutes, 95 degrees C for 30 seconds, primer annealing – 30 cycles at 57 degrees C for 30 seconds; DNA extension – 72 degrees C for 1 minutes; keep at 10 degrees C. After PCR amplification, we loaded our samples on to a gel; the samples were run at 90 volts for 30 minutes and the bands were visualized under UV light.

IV. Viability of *D. simulans* flies carrying the Mel mh allele

After confirming whether or not Mel mh was being expressed in the *D. simulans* genome, we wanted to find whether the *D. simulans* + Mel mh flies themselves were fully viable. To do this, we conducted controlled crosses between *D. simulans* + Mel mh males and females. From each stock of *D. simulans* + Mel mh flies (sim1029mh and sim1048mh) we selected either the homozygous flies or the heterozygous flies which we could differentiate between based on eye color (homozygous – dark red, heterozygous – light red). For a given homozygous cross, we mated 5 homozygous virgin female flies to 5 homozygous virgin male flies from the same stock.

The crosses were kept at room temperature and were flipped every 2-3 days for three weeks. Ten days after the first mating, we began to collect and count the progeny from these crosses. The progeny was scored three times a week for 3 weeks after the original mating date. We repeated the same cross scheme for the heterozygous flies. Both the homozygous and heterozygous crosses were run in triplicates for any given mating to account for the fact that each cross size was quite small. Based on our hypothesis that Mel mh in the *D. simulans* + Mel mh was not having any adverse effect on the viability of these transgenic flies, in the cross between homozygous flies, we expected to see approximately equal numbers of male and female progeny between each *D. simulans* + Mel mh strain and their *D. simulans* counterpart. We used a t-test to see if the difference in progeny between the two crosses was statistically significant or not. In heterozygous viability crosses, we expected that the progeny of these fly crosses would produce in the appropriate ratio of phenotypes: ¼ white eyed flies – offspring who received 0 copies of Mel mh from either the father or the mother, ½ light orange eyed – offspring who received 1 copy of Mel mh from either the father or the mother and ¼ dark orange eyes – offspring who received 2 copies of Mel mh, one from the mother and one from the father. We used a c Chi Square Goodness of Fit test to see if our data matched with the expected ratio of phenotypes.

V. Testing the role of Mel Mh in *D. simulans* x *D. melanogaster* hybrid cross

Our main objective was to test the candidate gene, *Maternal haploid* to determine whether in a hybrid cross between *D. simulans* and *D. melanogaster*, it contributes to female progeny viability. To do so, we crossed male *D. melanogaster* flies (from Canton-S strain) with virgin female flies from either sim1029 or sim1048, or sim1029mh and sim1048mh. We used the *D. simulans* x *D. melanogaster* crosses as controls for the *D. simulans* + Mel mh x *D. melanogaster* experimental crosses. Per crosses, we mated on average 15 females to 15 males in the parental cross and all crosses were maintained at room temperature. All of the female parents were virgins and the crosses were flipped every week for three weeks after the initial mating date. Over the span of 8 months, we attempted to create as many crosses and collect as much progeny as possible. With this data, we expected to see a larger number of females progeny from *D. simulans* + Mel mh x *D. melanogaster* than from the *D. simulans* x *D. melanogaster* progeny if our hypothesis that the presence of Mel mh contributes to increased female hybrid viability were true.

Results

The following tables and figures represent the results obtained from our experiments.

I. Expression of the *D. melanogaster* mh Allele in *D. simulans* + Mel mh

To ensure that Mel mh was in fact being expressed in the *D. simulans* genome, we conducted RT PCR experiments to convert RNA to cDNA and then analyzed the results under UV light. Each sample was run in concurrence with a –RT sample to account for any contamination. Furthermore, each set of samples was run with either *D. melanogaster* designed primers or *D. simulans* designed primers. The images below show the results.

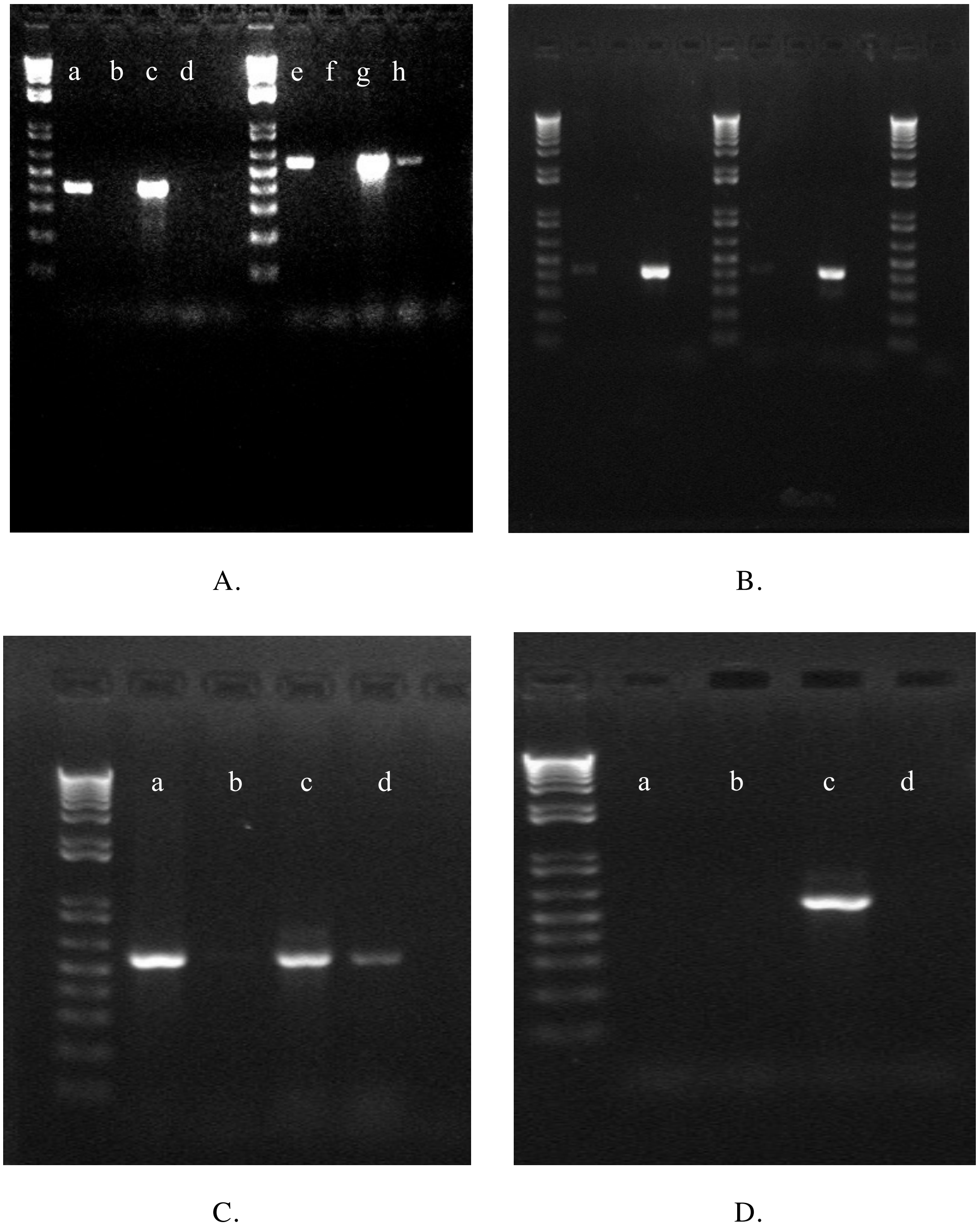


Figure 1 shows 4 PCR images. **A.** This gel corresponds to the cDNA obtained from the ovaries of Canton-S female flies. The gel is divided into two sections by the ladder in the center. The bands to the left of the ladder (lanes a-d) shows the amplified cDNA which had *D. melanogaster* primers. The bands in lanes a and c are the bands which had the reverse transcriptase enzyme while the empty lanes next to them (b and d) did not. The bands to the right of the ladder (lanes e-h) show the amplified cDNA which had *D. simulans* primers. The bands in lanes e and g are the bands which had the reverse transcriptase enzyme while the empty lanes next to them (f and h) did not. The band seen in lane h likely is the result of some cDNA contamination because

without the RT enzyme, no cDNA should have been amplified. We chose to omit this data. **B.** This gel shows the amplified cDNA of sim1029, sim1029mh, sim1048 and sim1048 (read left to right) with *D. melanogaster* primers. This gel shows that the *D. melanogaster* primers were able to anneal to all of the samples. **C**. This gel shows cDNA samples which were amplified at 55 degrees C in the thermocycler. The band to the very far left (lane a) is from the sim1029 cDNA + RT sample. This indicates that the duplicate sequence of *D. simulans* for which we chose primers was found in the sim1029 genome as well. The two bands seen directly to the right of that band (lane c and d) correspond to sim1029mh +RT and sim1029mh –RT run with *D. simulans* primers. We see that there is a band in the sim1029mh –RT lane meaning that some contamination has occurred. We chose to omit this data. **D.** This gel shows the sim1048 (lane a) and sim1048mh data. The band in lane c is the sim1048mh band. *D. simulans* primers were used.

After identifying which bands had appropriately amplified with either D. melanogaster and *D. simulans* primers, we performed a restriction digest on the bands to see if one or both duplicates of Sim mh are expressed in the genome of *D. simulans* flies and if Mel mh is expressed in *D. simulans* + Mel mh flies.

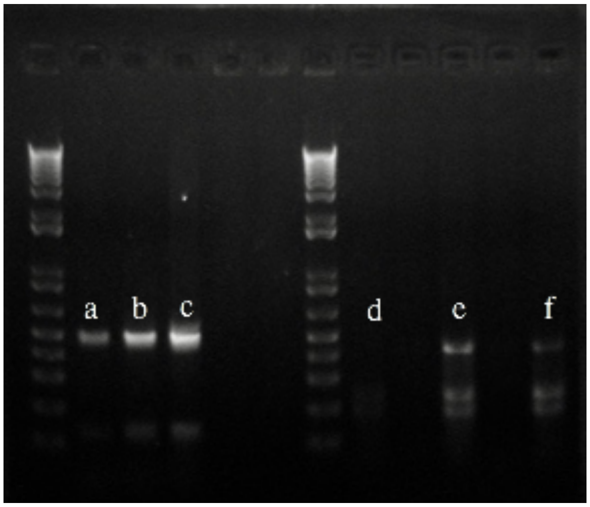


Figure 2 shows a restriction enzyme digest gel. Lanes (a), (b) and (c) are *D. melanogaster*, *D. simulans* + Mel mh and *D. simulans* respectively all with *D. simulans* primers. Bands show that one copy of *D. simulans* mh has been cut. (d), (e) and (f) are *D. melanogaster*, *D. simulans* + Mel mh and *D. simulans* + Mel mh respectively, with *D. melanogaster* primers. As expected, the lane a does not show any bands because *D. melanogaster* inherently has Mel mh. Conversely, the two *D. simulans* + Mel mh lanes (e and f) show that Mel mh is expressed in D. simulans + Mel mh.

II. Maternal Haploid in *D. simulans* and *D. melanogaster*

We wanted to find a method to quantify the divergence of Mel mh as well as of Sim mh_{Dup1} and Sim mh_{Dup2}. Further, we wanted to compare the dN/dS ratio of these genes to the dN/dS values of approximately 10,000 genes.

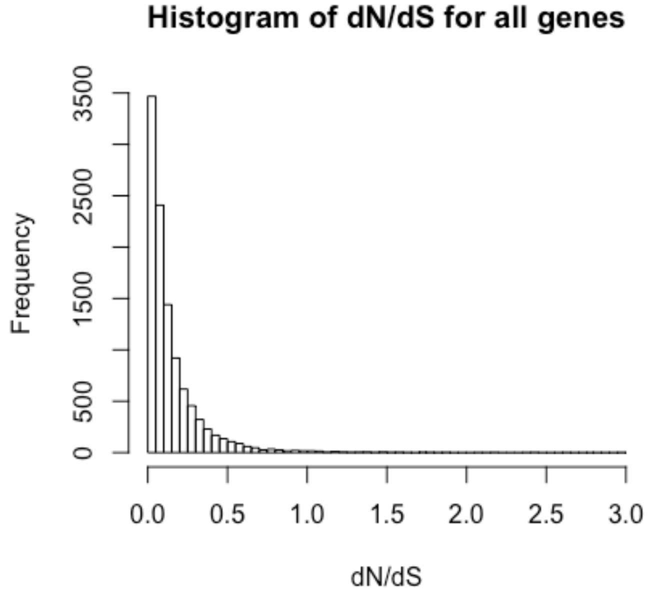


Figure 3 gives the number of genes (out of N = 10577) for a given published dN/dS values (from flyDIVaS database, N = 10577; outliers omitted: dN/dS > 3, incomplete values for dN or dS) in histogram form.

Since the majority of data falls within the dN/dS range of 0.0-0.5, the following density curve shows the same distribution as the histogram except in bell curve form. The values shown fall within the dN/dS ratio of 0.0-1.0. The lines on the curve correspond to the 25th, 50th and 75th percentile values of all of the data.

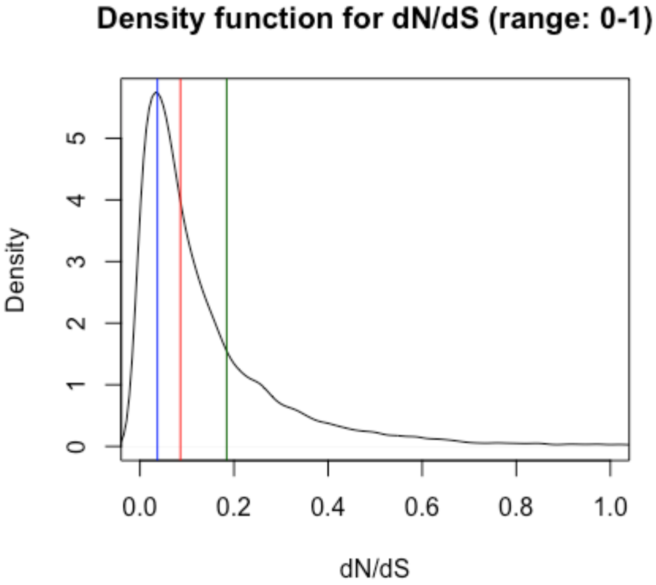


Figure 4 shows the histogram as a bell curve. The dN/dS values range from 0.0-1.0. The blue line is at the 25th percentile where dN/dS = 0.03840878, the red line is at the 50th percentile where dN/dS = 0.08740978 and the green line is at the 75th percentile where dN/dS = 0.1862334.

Given that we designed our own methods to find the dN/dS ratios for *Maternal haploid* in *D. melanogaster* and *D. simulans*, we wanted to verify that our methods were plausible. To do so, we repeated our methods on *Lhr*, a gene for which the dN/dS ratio has been documented accurately in the flyDIVaS database. The dN/dS value for *Lhr* in flyDIVaS is 0.6503207. After creating an alignment between the gene sequence of *Lhr* in *D. simulans* and the gene sequence of *Lhr* in *D. melanogaster*. After removing any intron sequences, we ran the alignment through SNAP which calculated the dN/dS ratio between the two genes to be 0.66168. The two values for dN/dS were nearly identical thus proving that our method for calculating dN/dS was valid.

By applying our methods to *Maternal haploid* we obtained the following results:

Genes	**dN/dS ratio**	**Percentile**
Mel mh & Sim mh_{Dup1}	0.50841	95.26th
Mel mh & Sim mh_{Dup2}	0.35532	90.28th
Sim mh_{Dup1} & Sim mh_{Dup2}	0.56459	96.27th

Table 1 gives the dN/dS ratios and percentiles for the genes listed in column 1. The larger the dN/dS ratio, the more divergent the genes.

We can look at where these values fall among the overall distribution of dN/dS values for the N>10,000 genes from the flyDIVaS database. Using the histogram above (**Figure 3**), we can plot these values at their respective percentiles.

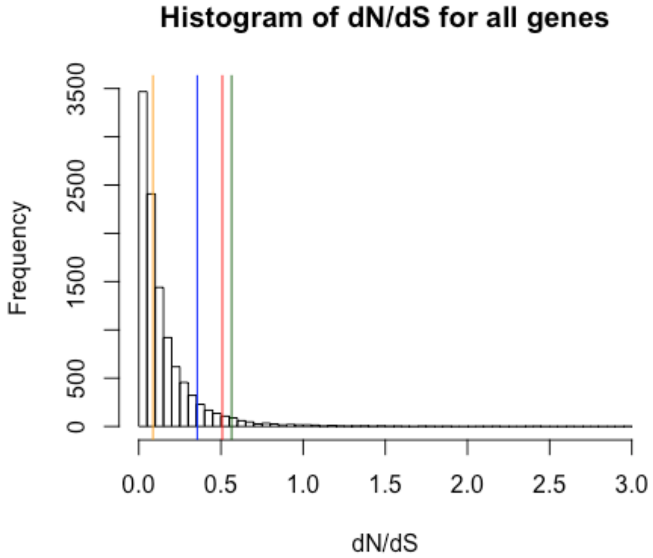


Figure 5 gives the number of genes (out of N = 10577) for a given published dN/dS values (from flyDIVaS database, N = 10577; outliers omitted: dN/dS > 3, incomplete values for dN or dS) in histogram form. The yellow line shows the median dN/dS value for all the genes given in the flyDIVaS database (dN/dS = 0.08740978). The blue line corresponds to the dN/dS ratio between Mel mh & Sim mh_{Dup2} which is at the 90.28th percentile. The red line corresponds to the dN/dS ratio between Mel mh & Sim mh_{Dup1} which is at the 95.26th percentile and the green line corresponds to the dN/dS ratio between Sim mh_{Dup1} & Sim mh_{Dup2} which is at the 96.27th percentile.

III. Function of Mel Mh in *D. simulans* x *D. melanogaster* hybrid cross

Over the span of several weeks, we collected progeny from the control hybrid cross (*D. simulans* x *D. melanogaster*) as well as the experimental hybrid cross (*D. Simulans* + Mel mh x *D. melanogaster*). We compared the percentage of female progeny from the control hybrid cross to the experimental hybrid cross from both the 1029 strain of *D. simulans* female parents and 1048 strain of *D. simulans* female parents.

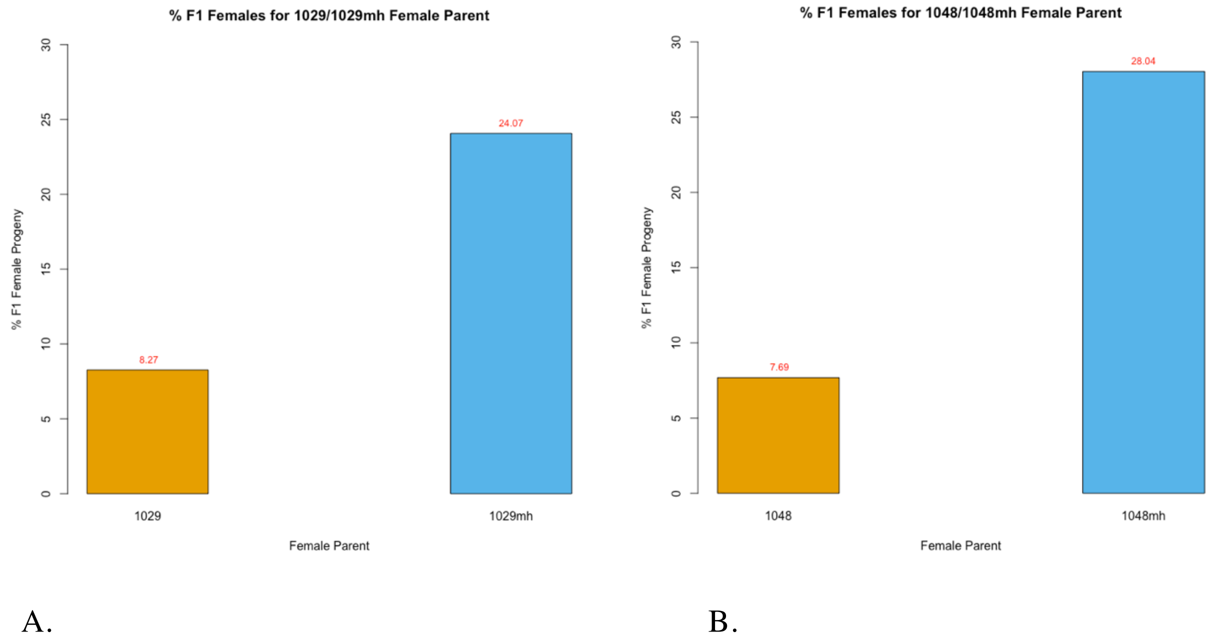


Figure 6 gives two bar plots which show the variation in female hybrid progeny percentage (percentage is given as the red number above the bar) with and without the presence of Mel mh in the *D. simulans* mother's genome. The orange bars correspond to the control cross (*D. melanogaster* males x *D. simulans* females) and the blue bars correspond to the experimental cross (*D. melanogaster* males x *D. simulans* + Mel mh females). **A.** refers to the comparison between female hybrid progeny from 1029 and 1029mh mothers and **B.** refers to progeny from 1048 versus 1048mh mothers.

Once the progeny had been collected, we wanted to test whether or not there was a difference between progeny counts from the hybrid crosses with normal *D. simulans* mothers versus *D. simulans* + Mel mh mothers. We performed a Chi Square test which compared the number of females and male progeny from the control cross to the number of female and male progeny from the experimental cross. We found that when comparing counts between 1029 and 1029mh, the p-value was less than 2.2 x 10^{-16} and when comparing between 1048 and 1048mh, the p-value was equal to 1.313 x 10^{-10}. Both of these values are below our set p-value of 0.05 and thus, we rejected the null hypothesis and determined that our data was statistically significant. The presence of Mel mh in the *D. simulans* genome does in fact increase hybrid female progeny viability.

IV. Additional Statistical Tests

To obtain a sufficient number of hybrid offspring, we conducted nearly 80 crosses between *D. simulans* x *D. melanogaster* flies and *D. simulans* + Mel mh x *D. melanogaster* flies over the course of 8 months. While we did our best to control for environmental factors such as temperature, mating flies in New York could have caused fluctuation in our results. Even looking at the raw data, it is clear that the flies produced slowly in the early months of 2019 whereas in July, the crosses were much more successful, and more progeny were produced. In order to account for this variability in temperature and other factors, we decided to perform another statistical test on the data. This time however, rather than pooling all of the data, we chose specific dates in the data on which both a *D. simulans* x *D. melanogaster* cross was performed and a *D. simulans* + Mel mh x *D. melanogaster* cross was done of the same strain. In other words, if a sim1029 was mated with Canton-S, then a sim1029mh was mated with Canton-S on the same day. Selecting for data in such a way narrowed our sample pool to exactly 10 data points for each pair (control *D. simulans* and experimental *D. simulans* + Mel mh) of crosses. We then computed the percentage of female progeny collected on a given day from each cross in the pair and performed a Wilcox test on all of the data to determine whether the paired data was still plausible while controlling for external factors. Our p-value when comparing the data between 1029 and 1029mh mothers was 0.03599 and between 1048 and 1048mh mothers was 0.007289. Both of these values are less than the set p-value of 0.05 and thus our initial hypothesis that Mel mh has an effect on female hybrid viability still holds true.

In our experiments, the presence of Mel mh in a *D. simulans* transgenic fly was identified by the orange eye color of the fly as opposed to the white eye color observed in normal *D. simulans* flies. If a fly had dark orange eye color, it had two copies of Mel mh in its genome (one from the father and one from the mother) and if the progeny had a light orange eye color it only had one copy of the Mel mh. We conducted controlled viability crosses on *D. simulans* flies and on D. simulans transgenic flies to see if the presence of any amount of Mel mh had an adverse effect on progeny rates. These viability tests crossed *D. simulans* flies to themselves and then crossed homozygous (two copies of Mel mh) *D. simulans* + Mel mh flies to themselves. Based on the t-test performed, we found that there was no statistically significant difference between the number of males and female progeny from a given cross of *D. simulans* x *D. simulans* or *D. simulans* + Mel mh x *D. simulans* + Mel mh. The p-value among the comparison of progeny rates from 1029 and 1029mh mothers was 0.278 and between 1048 and 1048mh was 0.344. Both values are greater than 0.05 and thus there is no statistical difference between males and females in homozygous crosses. Likewise, in heterozygous viability crosses, we wanted to ensure that the progeny appeared in the appropriate ratio of phenotypes: ¼ white eyed (no copies of Mel mh), ½ light orange eyed (1 copy of Mel mh) and ¼ dark orange eyes (2 copies of Mel mh). To test for statistical significance between the obtained ratio of phenotypes from the data versus the expected ratio, we performed a Chi Square Goodness of Fit test. We found that among progeny from the 1029mh transgenic mother, the p-value was 0.07217 and with progeny from the 1048mh mother, the p-value was 0.1161. In both cases, the p-value was more than 0.05. Thus, we conclude that there is no statistical difference between expected ratio of phenotypes and actual data.

Discussion

We have shown that the presence of Mel mh in the *D. simulans* genome increases female progeny viability when the hybrid species is created. We first demonstrated that Mel mh was expressed in a *D. simulans* + Mel mh transgenic fly (**Figure 2**) and concluded that the presence of Mel mh in a *D. simulans* + Mel mh transgenic fly did not impact the viability of the cross in comparison to the viability of a normal *D. simulans* cross. We then quantified the divergence between Mel mh and the Sim mh duplicate orthologs by comparing their dN/dS scores and determined that Mel mh is more diverged from Sim mh than the two duplicates are from one another (**Table 1**). Additionally, we found that *Mh* is a gene which is highly divergent compared to the majority of *Drosophila* genes as is shown by the fact that the dN/dS values for *Maternal haploid* fall above the 90th percentile (**Figure 5**). Finally, we found that compared to the percentage of female progeny from a control cross using a normal *D. simulans* female parent, the percentage of female progeny from an experimental cross using *D. simulans* + Mel mh transgenic females was significantly greater (**Figure 6**).

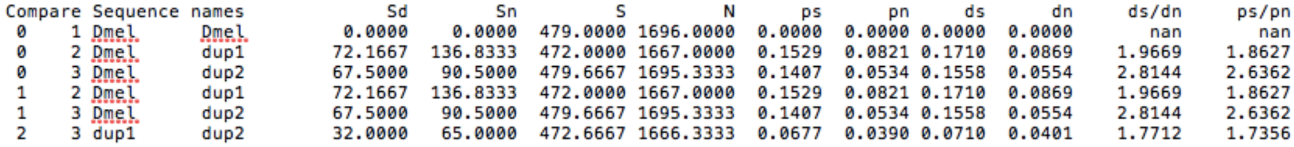
We hypothesized that the presence of Mel mh in the *D. simulans* genome would allow for increased female viability among the progeny of a hybrid cross. While our data shows this to be true among both strains of flies used in this experiment (**Figure 6**), there are a few other aspects of the experiment to be considered. For instance, hybrid matings can be quite difficult and in our case, from any one given cross, even obtaining 20 progeny was difficult. Often, the virgin female *D. simulans* + Mel mh flies would die before mating with the Canton-S males. These virgin female flies are yellow bodied and white eyed. Such flies tend to be mating deficient. One way we thought to alleviate this issue was to potentially use *D. simulans* flies which were yellow minus in the hope that they would mate more readily with Canton-S. However, eventually our original cross schemes began to produce a sizeable amount of progeny. To obtain enough hybrid offspring, we conducted nearly 80 crosses between *D. simulans* and *D. melanogaster* flies and *D. simulans* + Mel mh and *D. melanogaster* flies over the course of 8 months. We were able to collect nearly 200 flies per cross (800 flies total) which is still an appreciable sample size. Thus, we were confident in our result that Mel mh did in fact have an effect on increasing hybrid female progeny viability.

In addition, given that these matings occurred over the span of several months, with the changing weather patterns of upstate New York, we hypothesized that a weather-related factor may have had an effect on the data. Further, just by looking at the raw data, it is clear that the flies produced slowly in the early months of 2019 whereas in July 2019, the crosses were much more successful. Thus, we wanted to ensure that the date on which a certain mating occurred did not have any effect on the number of offspring produced from the cross. While our original test on the data took aggregate values of progeny from each cross and compared them using a Chi Square test, pairing the data based on date seemed like a more accurate control of temperature and other environmental factors. Specific dates were chosen on which both a control and experimental cross were done of the same strain. This narrowed our sample pool to 10 data points for each pair (control *D. simulans* and experimental *D. simulans* + Mel mh) of crosses. The percentage of female progeny collected on a given day was computed from each cross in the pair and Wilcox test was performed on the data. The test results determined that even by specifically looking at data on given dates across the span of the collection period, the p-value for both types of D. simulans flies was less than 0.05. Therefore, the data is still statistically significant and regardless of time of year and temperature fluctuations, our initial hypothesis that Mel mh has an effect on female hybrid viability is still supported.

When the Mel mh gene was inserted into the *D. simulans* genome, it contained a wild type eye marker which was used to identify the presence of Mel mh in the transgenic flies. When these transgenic flies mated with each other to produce progeny from which virgin females were collected as female parents for the experimental cross, the offspring may have been either homozygous for Mel mh, heterozygous for Mel mh or neither. Accounting for the number of copies of Mel mh in the female parent fly is important because by mixing homozygotes and heterozygotes, we could be impacting the likelihood of Mel mh to be passed down by the mother (i.e. if she has one copy she has less of a chance than if she has two copies) to her offspring. Given that the mother also passes down Sim mh and we are still unsure about the exact functionality of the gene, we can be more certain about the effects of Mel mh if we were to use only homozygous females. However, we chose to pool the homozygote and heterozygote mothers because regardless of how many copies of Mel mh she had, only the mother could pass along Mh since it is a maternal gene.

To alleviate some of the concern associated with using homozygous and heterozygous female parents, we conducted viability crosses between *D. simulans* flies to ensure that with the presence of either 2 copies (homozygous) or 1 copy (heterozygous) of Mel mh, the transgenic flies still produced viable progeny. For the homozygous experiment, we crossed *D. simulans* flies to themselves and then crossed homozygous (two copies of Mel mh) *D. simulans* + Mel mh flies to themselves. Based on the T-test results, we found that there was no statistically significant difference between the number of males and female progeny from a given cross of *D. simulans* x *D. simulans* or *D. simulans* + Mel mh x *D. simulans* + Mel mh. The heterozygous flies were expected to produce progeny in a particular ratio of ¼: ½: ¼ based on eye color phenotype. The proportions within the progeny collected from our heterozygous crosses matched these expected proportions of phenotypes. The Chi Square Goodness of Fit test helped confirm this. While these crosses only tell us that regardless of heterozygous or homozygous female parents, the transgenic strain is still as viable as the normal *D. simulans* stock, at least we can confidently assume that within the hybrid crosses, the presence of one versus two copies of Mel mh should not have an effect on the number of progeny produced.

Furthermore, our results from **Table 1** show that the dN/dS ratio for Mel mh and both duplicates of Sim mh fall above the 90th percentile in comparison to several other *Drosophila* genes. This leads us to believe that knowing that Mel mh and Sim mh are quite divergent from one another and to assume that the mother would pass along Mel mh if she had it, regardless of how many copies were in her genome. A closer look at the data shows that between Sim mh_{Dup1} & Sim mh_{Dup2} there is a larger dN/dS ratio than between Mel mh & Sim mh_{Dup1} and between Mel mh & Sim mh_{Dup2}. Intuitively, this does not make much sense because given that duplicate copies of genes are usually quite similar but differ in functionality, we would assume that Sim mh_{Dup1} & Sim mh_{Dup2} would be less diverged from one another than Mel mh and either duplicate of Sim mh.



```
Compare Sequence names             Sd        Sn         S         N       ps       pn     ds      dn       ds/dn       ps/pn
 0     1 Dmel      Dmel        0.0000    0.0000  479.0000 1696.0000  0.0000  0.0000 0.0000  0.0000         nan         nan
 0     2 Dmel      dup1       72.1667  136.8333  472.0000 1667.0000  0.1529  0.0821 0.1710  0.0869      1.9669      1.8627
 0     3 Dmel      dup2       67.5000   90.5000  479.6667 1695.3333  0.1407  0.0534 0.1558  0.0554      2.8144      2.6362
 1     2 Dmel      dup1       72.1667  136.8333  472.0000 1667.0000  0.1529  0.0821 0.1710  0.0869      1.9669      1.8627
 1     3 Dmel      dup2       67.5000   90.5000  479.6667 1695.3333  0.1407  0.0534 0.1558  0.0554      2.8144      2.6362
 2     3 dup1      dup2       32.0000   65.0000  472.6667 1666.3333  0.0677  0.0390 0.0710  0.0401      1.7712      1.7356
```

Figure 7 shows the output data from SNAP which caculates the dN/dS ratio for genes.

The above table is the output from SNAP when the dN/dS values were calculated from the alignment of Mel mh, Sim mh_{Dup1} and Sim $\text{mh}_{\text{Dup2.}}$ Looking at the dN column specifically, we see that between Sim mh_{Dup1} and Sim mh_{Dup2} (last row) the value is quite small, especially in comparison to the previous rows which compare Sim mh to Mel mh. Since dN refers to the nonsynonymous substitutions between two gene sequences. Thus, the smaller the number, the less variation there is between sequences.

The dN/dS value gives us some indication of the molecular evolution of these genes. A dN/dS ratio of less than 1 implies purifying selection – a process by which deleterious alleles are selected against. If the ratio is equal to 1, no selection for mutant alleles is occurring and further, a ratio greater than 1 indicates that the mutant alleles are causing natural selection to occur and thus driving molecular change and evolution. In our case, while our dN/dS values fall above the 90th percentile in comparison to several other *Drosophila* genes, our ratios also tell us that the genes are likely involved in the process of purifying selection by which harmful alleles are negatively selected for.

Ultimately, the purpose of our experiments were to know if *Maternal haploid* could be a gene which contributes to female hybrid viability in crosses between *D. melanogaster* and *D. simulans* flies. Our data show that it does in fact serve in that function and more over we were able to support this reasoning by showing that Mel mh is expressed in the *D. simulans* genome. This study has given us more insight into how hybrid incompatibilities function and how a gene like *Maternal haploid* can be vital for species survival. In addition, quantifying the divergence of *Maternal haploid* allows us to make conclusions about the way in which particular genes evolve which continues to help piece together how, in the broader context, species evolve.

Acknowledgements

I give my sincere thanks to Dr. Daniel Barbash for the immense support he has given me since 2016. For his continued faith in me to succeed, beginning from my first week as an undergraduate student at Cornell University, I am indebted to Dr. Barbash for his kindness and mentorship. I am grateful to him for his meticulous guidance as I wrote this thesis in pursuit of Honors from the Biological Sciences Program.

I extend my deepest appreciation to Dean Castillo, who through his patience and guidance, has been a wonderful mentor to me. I am truly lucky to have worked under his supervision and to have gained confidence in my own skills under his uplifting tutelage. To my first mentor in the lab, Shuqing Ji, who introduced me to the field of Drosophila genetics, I thank you for your encouragement and advice. Finally, to all of those, past and present who have been a part of Dr. Barbash's lab and who have been a constant source of support these past few years, I am greatly appreciative.

I thank the Thesis Committee from the Ecology and Evolutionary Biology department of the Biological Sciences program for their willingness to review my work. I also thank my undergraduate advisor Dr. William Crepet, the College of Agriculture and Life Sciences, and Cornell University itself for the opportunity to pursue my Bachelor's in Science degree at such a renowned university.

To my friends, who have sat with me through several late nights in the lab, been a source of motivation as I wrote my thesis and have always encouraged me to pursue my goals, thank you.

Finally, I thank my parents, Dr. Natesan Venkateswaran, Jayapreetha Natesan and my younger brother Sanjay for their utmost love and support as I have navigated my undergraduate years. From answering every phone call regardless of time of day, to offering and providing help in every situation, my gratitude transcends beyond limits.

As my time at Cornell University comes to a close, I will cherish the memorable journey it has been and will carry the lessons I have learned with me through the rest of my life.

Literature Cited

Assis, Raquel. “Drosophila duplicate genes evolve new functions on the fly.” *Fly* vol. 8,2 (2014)

Barbash, D.A. Ninety years of *Drosophila melanogaster* hybrids. Genetics *186*, 1–8 (2010)

Brideau, N. J., et al.Two Dobzhansky-Muller genes interact to cause hybrid lethality in *Drosophila*. (2006)

Ferree PM, Barbash DA Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. (2009)

Georgia Institute of Technology. "Seeing double: New study explains evolution of duplicate genes." ScienceDaily. ScienceDaily, 7 April (2014)

Johnson, N. A. Gene interaction and the origin of species. *Epistasis and the Evolutionary Process*,197–212 New York, Oxford University (2000)

Johnson, N. Hybrid incompatibility and speciation. *Nature Education*1(1):20 (2008)

Sawamura K, Watanabe T. K, Yamamoto M-T Hybrid lethal systems in the *Drosophila melanogaster* species complex. (1993)

Stern DL, Crocker J, Ding Y, Frankel N, Kappes G, Kim E, Kuzmickas R, Lemire A, Mast JD, and Picard S. Genetic and transgenic reagents for *Drosophila simulans, D. mauritiana, D. yakuba, D. santomea, and D. virilis*. (2017)

Tang X, et al. Maternal Haploid, a Metalloprotease Enriched at the Largest Satellite Repeat and Essential for Genome Integrity in *Drosophila* Embryos (2017)