Molecular Population Genetics of One Candidate Gene for Conspecific Sperm Precedence in *Drosophila*

Ashley Morris

Supervisor: Dr. A. Civetta

A thesis submitted to the Department of Biology, University of Winnipeg,

515 Portage Ave., Winnipeg, Manitoba R3B 2E9

Abstract

In several organisms, including *Drosophila*, the success of a male's sperm in multiple inseminated females depends on the male's genotype. In conspecific crosses, the last male to mate always fathers more than half of the progeny. However, when a female mates with a conspecific and a heterospecific male, the majority of the progeny is fathered by the male of the same species. This phenomenon is known as conspecific sperm precedence (CSP) and is a common post-mating pre-zygotic isolation barrier that occurs between closely related species. Prior work on the genetic basis of CSP between *D. simulans* and *D. mauritiana* identified fifteen candidate genes that had significant differences in either gene expression or DNA sequence composition between species. A third of the gene candidates were located within chromosomal map position 89B. My research focused on using a molecular population genetics analytical approach to try to validate or reject the candidate gene CG14891 as an agent promoting CSP. DNA sequence data of 8 strains of *D. simulans* and 1 strain of *D. mauritiana* for CG14891 were analyzed by comparing polymorphism and divergence data. The low rate of polymorphisms found within this testes-expressed gene suggests it may be important for the reproductive ability of the species. Information gathered from tests using both intraspecific and interspecies nucleotide changes suggest a trend towards a larger accumulation of non-synonymous changes between species than within species. This is expected under adaptive diversification between species. Overall, the signals for selection are weak, which is not unexpected in comparisons using closely related species. Thus, it will be worth expanding this kind of analysis to other previously identified candidate genes, especially within the 89B position.

Acknowledgments

I would like to thank my supervisor, Dr. A. Civetta, for this opportunity and for all of the advice and guidance throughout the year. I would also like to thank my committee members, Dr. Franck, and Dr. Good-Avila for their support. I would also like to extend thanks to Dr. Wiegand for keeping me informed and offering suggestions. Lastly, I would like to thank Barb and all of the members of the lab for their help along the way.

Table of Contents

Abstract ii

Acknowledgments iii

Table of Contents iv

List of Tables vi

List of Figures viii

List of Appendices ix

1.0 Introduction 1

1.1-The *Drosophila melanogaster* Subgroup 1

1.2-Sexual Selection and Sperm Competition 2

1.3-Speciation 4

1.4-Conspecific Male Precedence 5

1.5-Candidate Genes for Conspecific Sperm Precedence 6

1.6-Population Genetics 8

1.7-DNA Sequencing 9

1.8-Objectives 10

1.9-Hypothesis 10

2.0 Materials and Methods 11

2.1-*Drosophila* Stocks 11

2.2-DNA Extraction 12

2.3-Polymerase Chain Reaction Amplification 13

2.4-Confirmation of PCR Product 14

2.5-Cleaning of PCR Product 15

2.6-Quantification and Sequencing Reaction 16

2.7-DNA Precipitation 18

2.8-Sequence Data Analyses 19

3.0 Results 20

3.1-Sequence Data Information 20

3.2-Within Species Polymorphism 21

3.3-Between Species Divergence 23

4.0 Discussion 28

5.0 Conclusions 36

6.0 References 37

7.0 Appendices 42

List of Tables

Table 1. *Drosophila* standard cornmeal-molasses (CM) medium.

Table 2: DNA extraction buffer protocol.

Table 3: Protocol for PCR amplification.

Table 4: Protocol used to prepare 10 X TBE buffer.

Table 5: The reference chart used to calculate the volume of cleaned PCR product needed for the sequencing reaction.

Table 6: The protocol for the sequencing reaction.

Table 7: The solution used to stop the sequencing reaction.

Table 8: Summary of the total amount of analyzable nucleotides for the candidate gene CG14891 for eight strains of *D. simulans* and one strain of *D. mauritiana*.

Table 9: Summary of the results of statistical tests based on polymorphism within *D. simulans* and divergence between *D. simulans* and *D. mauritiana*.

Table 10: The McDonald and Kreitman Test for the coding region of the candidate gene CG14891.

Table 11: The ratio of non-synonymous to synonymous polymorphisms calculated for *D. simulans* and the ratio of non-synonymous to synonymous nucleotide changes between species.

Table 12: Estimates of nucleotide heterozygosity for 15 genes located on the third chromosome of *D. simulans*.

List of Figures

Figure 1: The consensus phylogeny of the nine species of *Drosophila* in the *D. melanogaster* subgroup.

Figure 2: Sliding window analysis of the ratio of non-synonymous to synonymous polymorphism within *D. simulans* and ratio of non-synonymous to synonymous divergence between *D. simulans* and *D. mauritiana*.

List of Appendices

Appendix 1: Forward and reverse primer sequences used during polymerase chain reaction and sequencing reaction.

Appendix 2: A list of non-synonymous codon changes in the analyzable region of the candidate gene CG14891 between eight strains of *D. simulans.*

Appendix 3: A list of non-synonymous codon changes in the analyzable region of the candidate gene CG14891 between the eight strains of *D. simulans* and the single strain of *D.* mauritiana.

1.0 Introduction

1.1 The *Drosophila melanogaster* Subgroup

The *Drosophila melanogaster* subgroup is composed of nine species, with the species *D.melanogaster, D. simulans, D. mauritiana,* and *D. sechellia* forming the *D. melanogaster* complex (Figure 1). The species *D. melanogaster* is the most ancient member of the *D. melanogaster* complex and has an estimated divergence time of three million years from the other three more closely related sibling species known as the *simulans* clade (*D. simulans, D. mauritiana* and *D. sechellia*) (Kilman et al. 2000), (Figure 1). *D. simulans* has a cosmopolitan distribution whereas *D. mauritiana* is endemic to the Mauritius Islands and *D. sechellia* to the Seychelles islands; both island groups are located in southern Africa (David et al. 2007).

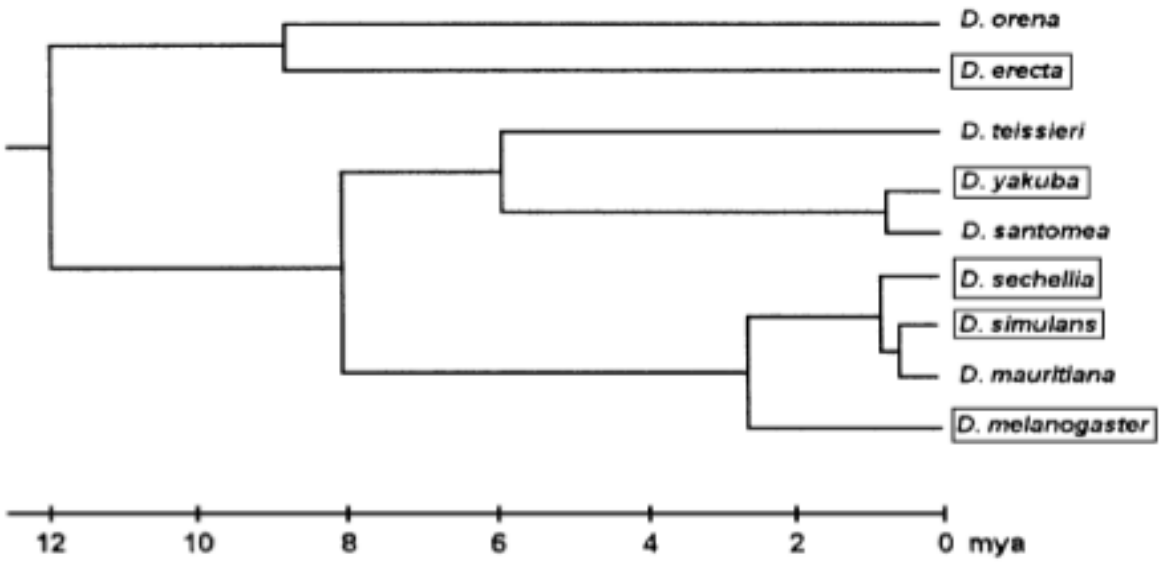

Figure 1: The consensus phylogeny of the nine species of *Drosophila* in the *D. melanogaster* subgroup. *D. simulans* and *D. mauritiana* are the most closely related species within the *simulans* clade (David et al. 2007). The boxes indicate the species in which the entire genome has been sequenced.

The divergence times of these species are continually under investigation, however no consensus divergence time has been determined (David et al. 2007). One experiment involved comparative analyses of DNA sequence variation within and between species. Fourteen different genes were used to delimit the evolutionary origins and divergence between species of the *simulans* clade (Kilman et al. 2000). Based on these results, Kilman et al. (2000) estimated that *D. simulans* and *D. sechellia* diverged into separate species approximately 413,000 years ago while *D. simulans* and *D. mauritiana* diverged into separate species approximately 263,000 years ago. Overall, the genetic distance between *D. simulans* and *D. mauritiana* is consistently shown to be one of the lowest within the *D. melanogaster* subgroup (Ohnishi et al. 1983; Solignac and Monnerot, 1986). Despite their divergence time, all members of the *simulans* clade are very similar except for the male's behaviour and the genital arch. As well, all three species are fully capable of hybridizing and producing interspecies hybrids. Hybridization results in an F1 generation in which the females are fertile and the males are sterile (David et al. 2007).

1.2 Sexual Selection and Sperm Competition

Sexual reproduction allows for shuffling of genetic information between genomes, creating genetic variation that provides new solutions to environmental challenges. The ability to reproduce sexually results in increasing opportunities for survival; however successful reproduction requires the use of heavy investments. Examples are seen in the evolution of complex ornaments used by males to secure mating, which is the basis of the theory of sexual selection originally proposed by Charles Darwin. Darwin

proposed that complex traits found in males provide an advantage in terms of non-random differential reproductive success that arise from competition for access to mates (Singh et al. 2002). It is also known that as the production of gametes and mating requires a large energy contribution, organisms have evolved different mechanisms that aid in concentrating reproductive energies on matings with the highest chance of success (Snook and Markow, 2002). However, a male needs to do more than simply mate with a female to ensure reproductive success (Civetta and Singh, 1999) as competition can continue after mating. Sperm competition can occur when two different sources of sperm enter the female, such that the sperms compete with each other for resources from the female prior to the formation of a zygote. The ability for a single sperm to compete and get to an ovum more quickly than any other sperm will result in the male that formed that sperm to be favoured (Ramm et al. 2005). Laboratory studies show that when a *Drosophila* female is mated to two males in succession, the last male to mate always fathers the majority (> 65%) of the offspring (Simmons, 2001). Field studies also show that *Drosophila* females produce progeny that are on average fathered by two or three males (Harshman and Clark, 1998), indicating that sperm competition occurs in nature.

The genus *Drosophila* exhibits a great diversity in sperm morphology; for example, there is a greater variation in sperm length in *Drosophila* than all other animal taxa combined (Simmons, 2001; Pitnick, 1996). As well, female *Drosophila* have the ability to store sperm from many previous matings for an extended length of time (Civetta, 1999). There are three proposed mechanisms by which sperm competition can occur; sperm incapacitation, sperm selection, and sperm displacement (Simmons,

2001). Sperm incapacitation involves the active inhibition of the sperm from the female's prior matings by a subsequent male's ejaculate. This is theorized to be accomplished by physiological adaptations in the male seminal fluid proteins or possibly morphological adaptations in the sperm of the last copulating male (Price et al. 1999). However, recent studies have raised some questions about the validity of the incapacitation hypothesis (Civetta et al. 2008; Holman, 2009). Sperm selection is defined as the ability of the female, at the gametic level, to specifically choose which sperm will fertilize an egg (Birkhead, 1998). This rather unusual ability is controversial. For example, studies on sperm viability in female storage have found no female-directed effect (Civetta et al. 2008; Holman, 2009). However, females are thought to have at least partial control over the utilization of sperm within their bodies; and recent studies have shown that female muscle contractions might influence sperm storage and dumping (Adams and Wolfner, 2007). The third method, sperm displacement, involves the displacement of the previous male's sperm to allow the sperm of the new male to not have to compete with previous sperm to fertilize the ova. This can be accomplished by sperm removal or sperm flushing. Sperm removal involves the direct removal of rival sperm from the female's sperm storage organ just prior to ejaculation; this is aided by adaptations of the male's genitalia (Simmons, 2001). Sperm flushing involves the indirect removal of the rival sperm from the female's sperm storage organ by the ejaculate of the copulating male which forces the previous sperm out (Simmons, 2001).

1.3 Speciation

The biological species concept equates speciation with the evolution of

reproductive isolation between populations (Chang, 2004). Reproductive isolation can occur as a side product of changes in reproductive traits which can be propelled by sexual selection. There are three main stages at which sexual isolation can occur; prior to mating, post-mating but prior to fertilization, or post-fertilization. Habitat isolation, mate discrimination, and differences in physical features separate species based on pre-mating factors. Hybrid sterility or inviability and zygote mortality pertain to reproductive isolation which occurs post-fertilization (Chang, 2004). Non-competitive and competitive forms of isolation occur post mating but pre-zygotically (Chang, 2004). Non-competitive forms of isolation include shortened duration of heterospecific copulation, decreased efficiency of heterospecific sperm transfer and use, and the reduced storage of heterospecific sperm (Chang, 2004). For example, in crosses between *D. simulans* males and *D. mauritiana* females, fewer hybrid offspring (mean = 32.5 progeny) are produced when compared to matings between conspecific *D. simulans* (mean =161.7 progeny) (Price et al. 2001). However, in crosses between *D. simulans* females and *D. mauritiana* males, more offspring (mean = 125.9 progeny) was produced as a result of the heterospecific matings than the *D. mauritiana* conspecific matings (mean = 68.9 progeny) (Price et al. 2001). Competitive isolation includes the phenomenon of conspecific sperm precedence.

1.4 Conspecific Sperm Precedence

The evolution of reproductive isolation mechanisms have usually focused on events that occur either before copulation or after a zygote is formed (Price et al. 2000).

Conspecific sperm precedence (CSP), a form of competitive isolation, is a common post-mating, pre-zygotic isolation barrier that occurs between closely related species when the sperm from a heterospecific male and a conspecific male are present in a single female. Observations of CSP have been described in many species including beetles, crickets, and *Drosophila*. While crosses between a female and two or more males of the same species lead to a common pattern of the last male fathering the majority of progeny (Simmons, 2001); when a female mates with a conspecific male followed by a heterospecific male, the majority of the progeny are fathered by the conspecific male (Price et al. 2000). Chang (2004) found that CSP is problematic to study as it is difficult to judge whether and how CSP affects reproductive compatibility between closely related species in nature. A major problem is that few species actually hybridize in nature. As well, the rate at which CSP evolves is unclear, thus other reproductive isolating barriers may evolve before CSP (Chang, 2004). However sperm competition clearly happens in natural populations of *Drosophila* and it might, if not directly, play a role in species differentiation by reinforcing existing barriers to reproduction (Price, 1997).

1.5 Candidate Genes for Conspecific Sperm Precedence

Until recently only two studies, one in *Drosophila* (Civetta et al. 2002) and the other in crickets (Britch et al. 2007), have attempted to map genes contributing to conspecific sperm precedence. Both studies were completed using associations between random molecular marker segregation and differences in phenotype. Using *D. simulans* and *D. sechellia* introgressed lines, Civetta et al. (2002) found significant

quantitative trait loci (QTL's) on the second and third chromosomes. This result explained a great deal of the variation that was found among the introgressed lines in the progeny scored, but only when using low stringency statistical thresholds (Civetta et al. 2002). They also suggested that the differences in sexual traits between species may be at least partially due to complex interactions among alleles (Civetta et al. 2002). Britch et al. (2007) used F2 backcross females of crosses between two species of crickets, *Allonemobius fasciatus* and *A. socius,* which were crossed with males of these species. F2 backcross females, which are heterozygous for all candidate markers, were produced by hybridizing *A. fasciatus* males and *A. socius* females and backcrossed F1 females with males from first-generation lab-reared parental lines (Britch et al. 2007). They found many unlinked markers that are associated with either the reducing or the enhancing of the success of paternity of the conspecific male (Britch et al. 2007).

In a recent study, the effect of third chromosome *D. mauritiana* genome introgressions in a *D. simulans* background on second male paternity success was assayed. The assays involved mating *D. simulans* ebony females to *D. simulans* ebony males followed by a set of wild-type males from 60 different introgression strains. The assay took advantage of the fact that body colour of the progeny can be used to assign paternity and that the introgressions were previously mapped (Tao et al. 2003). The study found that introgressions within two regions of the third chromosome were sufficient to cause a breakdown in second male paternity success and to produce a phenotype similar to that observed when a *D. mauritiana* male is second to mate. Candidate genes were identified within the introgressed regions by taking advantage of

gene annotations from the *D. melanogaster* genome project. This was done by selecting genes with a possible role in reproduction. A list of 81 candidate genes was narrowed down using two different approaches. One was to test for genes showing evidence of adaptive diversification in either the *D. simulans* or *D. mauritiana* lineage. Such analysis took advantage of sequence data available from the 12 *Drosophila* genome project (*Drosophila* 12 Genomes Consortium, 2007) and the use of phylogenetic analytical tools (PAML; Yang, 1997). The other approach was to test differential male reproductive tract gene expression between *D. simulans* and *D. mauritiana* for all 81 candidate genes using quantitative real time PCR. In total, fifteen candidate genes within the mapped region that have considerable differences in either gene expression or DNA sequence between species were identified. A full one-third of these genes were found within the 89B region of the third chromosome (Levesque et al. 2010). Further sequence analysis of one such gene, CG14891, is the focus of my thesis. The only prior information known regarding this gene is that it is a gene that is expressed in the testes (Levesque et al. 2010).

1.6 Population Genetics

While the prior study (Levesque et al. 2010) found patterns of adaptive diversification for CG14891, the use of a single sequence as a representative of the species in a phylogenetic analysis might mask important population genetics signals that can be relevant in trying to understand the role of selection during gene divergence. Population genetics concerns factors such as natural selection, genetic drift, mutation, and gene flow, that contribute to the evolution of a population (Halliburton, 2004). The

integrated theory of molecular population genetics involves the neutral theory of molecular evolution. The core ideas of the neutral theory are that genetic differences between species as well as genetic polymorphisms within species are the result of neutral mutations and random drift. Then, much molecular variation is considered to be selectively neutral and the standard allele frequencies are a result of genetic drift and not natural selection (Kimura and Crow, 1964; King and Jukes, 1969). Mutations in single nucleotides may lead to new alleles which are subject to genetic drift; through drift these new alleles have the opportunity to become more common in the population and may eventually become fixed or lost completely. Neutralists often accept the possibility of selection but only as a force for removing deleterious mutations, that is, negative selection. On the other side, selectionists argue that an advantageous phenotype will become more common in a population and thus any allele linked to the phenotype will increase in frequency. Both theories agree that the majority of mutations are negative. Selectionists add that positive selection might lead to the fixation of advantageous alleles in different populations. The combined analysis of polymorphisms and gene divergence at the DNA sequence level can shed light on the role played by random factors versus selective pressures.

1.7 DNA Sequencing

The first reported DNA sequencing results were accomplished using the Maxam and Gilbert DNA sequencing method (Gilbert and Maxam, 1973). The improved process of DNA sequencing did not come about until 1977 (Sanger et al. 1977). The original technology was technically complex, involved the use of hazardous chemicals and was

a relatively slow process (Gilbert and Maxam, 1973); thus was quickly replaced. Kreitman (1983) was the first to apply techniques of DNA sequencing to population genetics. He focused on the alcohol dehydrogenase locus of *D. melanogaster* and found that in the eleven copies he sequenced, all were different **(**Kreitman, 1983). In the early 1990's automated sequencing became available which allowed for large portions of DNA to be sequenced relatively quickly. This breakthrough led to a wide variety of organisms to have their complete genome sequenced. The recent completion of 12 *Drosophila* genome sequences allowed for the preliminary analysis of CG14891 gene region of *D. melanogaster* to be used for comparison with the sequence of *D. simulans*.

1.8 Objectives

- To determine if one of the candidate genes (CG14891), based on comparisons with the neutral theory of molecular evolution is the driving force behind conspecific sperm precedence. This is to be accomplished through DNA sequence data analysis.

- To test whether adaptive selection has driven gene divergence between species while keeping population polymorphism within species low or balanced.

1.9 Hypothesis

- CG14891 is a factor in regards to CSP, however not the only factor.

- Adaptive selection will drive gene divergence between species while keeping population polymorphism within species low or balanced.

2.0 Materials and Methods

2.1 *Drosophila* Stocks

Drosophila simulans flies were provided by Dr. Andrew G. Clark (Cornell University) and the *Drosophila mauritiana* flies were purchased from the Drosophila Species Stock Centre (UC San Diego). The strain of *D. mauritiana* that was used in analysis was stock number 14021-0241.01. Flies were kept on a cornmeal molasses media (Table 1).

Table 1. *Drosophila* standard cornmeal-molasses (CM) medium. Water is heated to a boil in a steel pot. The first three ingredients are combined and added to the boiling water. The mixture is stirred until the boiling point is reached again; it is then removed from heat and molasses is added. The mixture is then cooled to 65 ^{o}C in a water bath. Tegosept and propionic acid are added and then it is poured into vials or bottles with a peristaltic pump. Yields approximately one hundred, 28.5 x 95 mm vials or eighteen, 237 ml bottles.

Ingredient	Quantity
Cornmeal	65g
Brewers Yeast	13g
Agar	6.5g
Cold Water	170ml
Boiling Water	760 ml
Refiners Molasses	45.5 ml
10% Tegosept*	20 ml
99% Propionic Acid	5 ml

* 50 g methyl 4-hydroxybenzoate in 500ml 95% ethanol.

These stocks were maintained by transferring adults to bottles containing fresh media approximately every 11 days. Flies were stored at room temperature.

2.2 DNA Extraction

Carbon dioxide gas was used to immobilize the flies, allowing for five flies to be removed from the storage vial and inserted into a micro-centrifuge tube. Approximately 250 μl of an homogenizing buffer was added to each set of five flies; the flies were then ground and mixed with a small pestle. The ingredients for the homogenizing buffer are shown below (Table 2).

Table 2: The ingredients and their respective quantities needed in the preparation of DNA extraction buffer.

Ingredient	Quantity
Sterile Double Distilled Water (SDDW)	237.5 μl
1M Tris HCl	2.5 μl
0.05M EDTA	5 μl
2.5 M NaCl	2.5 μl
20 mg/ml Proteinase K	2.5 μl

The extract was incubated at room temperature for thirty minutes and heated at 95 ^{o}C for three minutes. DNA extractions were stored at -20 ^{o}C until needed.

2.3 Polymerase Chain Reaction (PCR) Amplification

A two microlitre aliquot of the DNA from the extraction was placed into a PCR tube. Twenty-three microlitres of a PCR master mix was then added to each tube (Table 3). Forward and reverse PCR primer sequences used can be viewed in the appendix (Appendix 1).

Table 3: The quantity and ingredients needed in the PCR reaction. Note: Taq DNA polymerase was added last. Both forward and reverse primers were diluted 10 fold from a 100 mM stock.

Ingredient	Quantity
10X Thermopol PCR Buffer	2.5 μl
Deoxynucleotide Triphospate (dNTPs)	0.8 μl
10 nM Forward Primer	0.8 μl
10 nM Reverse Primer	0.8 μl
SDDW	18.5 μl
Taq DNA Polymerase	0.2 μl

The PCR tubes containing the DNA extraction and the PCR master-mix were immediately placed into a thermocycler. A specific thermoprofile on the MJ thermocycler was created to produce the best results for the gene. For CG14891, this involved initially heating the sample to 95 ^{o}C for 2 minutes. The next three steps of the cycle were then repeated 30 times; this involved heating to 95 ^{o}C for 45 seconds, cooling to 58 ^{o}C and held for another 30 seconds. The samples were then heated to

72 ^{o}C for 1.33 minutes. After thirty cycles were completed, the samples were held at 72 ^{o}C for 15 minutes and then held at 10 ^{o}C for storage. Once PCR products were removed from the cycler, they were stored at -20 ^{o}C.

2.4 Confirmation of PCR Product

Confirmation of the correct product was determined by fractionation through an agarose gel. A 1% agarose gel was prepared by adding 0.75 g agarose to 75 mL of 1X tris(hydroxymethyl) aminomethane (Tris), boric acid and ethylenediaminetetraacetic acid (EDTA) [TBE] to an erlenmeyer flask. 1X TBE was diluted 10 fold from 10X TBE; the instructions for the production of 10X TBE are shown below in Table 4.

Table 4: Ingredients and their respective quantities for 10X TBE. Note: This was all placed in a large beaker and enough SDDW was added to make 1 litre. The 10X TBE was autoclaved before use.

Ingredient	Quantity
Tris	108 g
Boric Acid	55 g
EDTA	5.85 g

The agarose was then dissolved in the 1X TBE; this was done by heating the flask until boiling occurred and then the flask was removed from the heat source. The Erlenmeyer flask was cooled to 60 ^{o}C and then 3.75 μl of ethidium bromide was added. The flask was swirled and poured into a gel box, in which it solidified in approximately

twenty minutes. 1X TBE was poured over the gel in the gel box until it was slightly higher than the surface of the gel. A 4 μl aliquot of a DNA ladder was pipetted into one well in the gel; a mixture of 3.5 μl of the PCR product along with 1 μl of bromophenol blue tracking dye was added to each subsequent well. The gel was run at 120 volts for approximately 40 minutes, after which, it was viewed under ultraviolet light. A picture of the gel was taken using a Gel Doc system.

2.5 Cleaning of PCR Products

Cleaning was accomplished by the use of E.Z.N.A. Cycle-Pure Kit (Omega biotek) #D6493 VWR CA101318-906. This kit consisted of DNA wash buffer, Buffer CP, Elution buffer, HiBind® DNA columns, and 2 ml collection tubes. The DNA wash buffer was reconstituted by adding 1 part DNA wash buffer to 4 parts of 95% ethanol. The PCR products of each original strain were combined into a 1.5 ml microcentrifuge tube. The volume was recorded and 4 times this volume of Buffer CP was added to each tube. The samples were vortexed, transferred to the Hiband® DNA columns and centrifuged at 10,400 rpm for 1 minute. The flow-through was discarded and the mini-column inserted back into the collection tube. A 700 μl portion of the reconstituted DNA wash buffer was added to each column and the samples were again centrifuged at 10,400 rpm for 1 minute. The flow-through was again discarded and the mini-column reinserted into the collection tube. A 500 μl portion of the reconstituted DNA wash buffer was then added to each column and centrifuged at 10,400 rpm for 1 minute. The flow through was discarded and the mini columns reinserted into the collection tubes. The HiBind® DNA columns were then centrifuged at 13,700 rpm for 2 minutes. The DNA

columns were then transferred to micro-centrifuge tubes and 30-50 μl of elution buffer was added to the mini-column, dependent upon the intensity of the bands in the previous agarose gel. The samples were incubated at room temperature for 1 minute and then centrifuged at 13,700 rpm for 1 minute. The cleaned DNA was stored at -20 ^{o}C.

2.6 Quantification and Sequencing Reaction

The amount of purified PCR products was calculated by running the cleaned PCR reactions in a 1% agarose gel and comparing the intensity of the sample bands to those in a mass ladder. As the mass values of each band on the ladder are known, it was possible to calculate the mass of DNA per microlitre of sample, based on the intensity of the band. Thus, the volume of sample needed for the sequencing reaction can be determined based on comparison to the reference (Table 5).

Table 5: The reference chart used to calculate the volume of cleaned PCR product needed for the sequencing reaction. With cleaned PCR products, 25-100 fmol of template had to be added to the new PCR tubes. Note: This table used in conjunction with quantification of the cleaned PCR product with an agarose gel and a mass ladder.

Size of Template (kb)	ng for 25 fmol	ng for 100 fmol
0.5	8.1	33
1.0	16	65
2.0	33	130

In total, nine primers were used in the sequencing reaction; three of which were forward and the other six were reverse (Appendix 1). The other ingredients needed for the sequencing reaction are listed below (Table 6).

Table 6: Ingredients used in the sequencing reaction. A volume of water was added such that the total solution was 20 μl. *An aliquot of 0.8 μl of each primer was added to each different strain of DNA; such that there were nine separate reactions for every strain, each with a single, different primer.

Ingredient	Quantity
Dye terminator cycle sequencing (DTCS) Quick Start Master- mix	2 μl
Sequencing Buffer	1.5 μl
Pellet Paint	0.25 μl
Cleaned PCR Product	Based on Calculations
*Primer	0.8 μl

The tubes were then placed in the MJ thermocycler and run under the sequencing program until completion. The sequencing reaction consisted of the samples being heated to 95 ^{o}C for 45 seconds, cooled to 50 ^{o}C for 45 seconds, and heated slightly to 60 ^{o}C for 4 minutes. This cycle was repeated thirty-four times. On the final cycle, the samples remained at 60 ^{o}C for another 8 minutes, before being cooled to 12 ^{o}C. Just prior to the completion of the program a stop solution was prepared and added to the end sequencing reaction. The ingredients for the preparation of the stop solution are listed in Table 7.

Table 7: Quantity of ingredients used in the preparation of the stop solution which was administered immediately following the sequencing reaction. Note: Glycogen was added last to the solution.

Ingredient	Quantity
3M NaOAC	2 μl
0.5M EDTA	0.4 μl
SDDW	1.6 μl
Glycogen	1 μl

2.7 DNA Precipitation

Next, 60 μl of cold 95% ethanol was added to each sample and mixed. The samples were then centrifuged at 14,000 rpm for 15 minutes at 4 ^{o}C. The supernatant was removed and 100 μl of cold 70% ethanol was added to each pellet. These samples were then centrifuged at 14,000 rpm for 5 minutes at 4 ^{o}C. Again, the supernatant was removed, 100 μl of 70% ethanol was added to each pellet, and centrifuged at 14,000 rpm for 5 minutes at 4 ^{o}C. The supernatant was again removed. The micro-centrifuge tubes were then inverted and the pellet was left to dry for 30 minutes. After, the tubes were turned upright, and the pellet was left to dry for another 15 minutes. During this time, the Beckman Coulter CEQ 2000XL automated sequencer was turned on and set up by entering a sample data arrangement in the CEQ sample setup module. Once the samples were dry, 40 μl of the sample loading solution (CEQ Kit) was added to each pellet and left for 15 minutes. Once the DNA was resuspended, the entire sample was transferred into the corresponding wells of the sequencing sample tray. One drop of mineral oil was used to cover each sample. The separation buffer was then loaded into

the corresponding wells of the buffer tray. Both trays were then loaded into the sequencer and left to run overnight. The sequencing reactions were completed using both forward and reverse primers from two independent PCR products. These sequence reads from different primers were combined to a single consensus sequence for each strain. To assemble the different sequence reads I used the alignment program ClustalX 2.0.10 (Thompson et al. 1997). Trace files were analyzed manually in cases where overlapping sequence data did not agree.

2.8 Sequence Data Analyses

Final consensus sequences from all *D. simulans* strains and *D. mauritiana* were aligned using ClustalX 2.0.10 (Thompson et al. 1997); analysis of divergence and polymorphism was examined using the program DnaSP (version 5.0) (Rozas et al. 2003). This software program works by performing population genetics analyses of DNA polymorphism data. These analyses included measures of sequence variation for within species as well as between species. In addition, DnaSP performs several tests of neutrality such as Tajima (Tajima, 1989), Fu and Li (Fu and Li, 1993), and McDonald and Kreitman (McDonald and Kreitman, 1991). Protein sequence translations were scanned for presence of functional domains using the MotifScan tool found in Expasy (http://www.expasy.ch/tools/dna.html).

3.0 Results

3.1 Sequence Data Information

In regards to the candidate gene CG14891, ten strains of *Drosophila simulans* and four strains of *Drosophila mauritiana* were sequenced; two *D. simulans* strains and three *D. mauritiana* strains were excluded from analysis due to incomplete data. In total there were sequence data sets from 8 strains of *D. simulans* and 1 strain of *D. mauritiana* used in the analysis. The *D. simulans* coding portion of the gene is 1,479 base pairs long compared to 1,488bp for *D. melanogaster* (http://flybase.org/). The gene consists of one protein coding region (exon) and does not contain any introns. I obtained sequence data for 1,407 nucleotide sites. Due to missing data (e.g. a nucleotide position that could not be solved during sequencing and had to be coded as N) or the presence of indels (i.e. gain or loss of nucleotides in some strains/species relative to others) only 1,119 nucleotide sites were comparable among all *D. simulans* strains and *D. mauritiana* . Within this complete region, 1,035 nucleotides were identical or conserved among all sequences and 84 nucleotide sites were different. The majority of differences (77), were rare or low in frequency as they were found in a single sequence (singletons) with the remaining differences (7) being shared between two sequences.

Table 8: Summary of the total amount of analyzable nucleotides for the candidate gene CG14891 for eight strains of *D. simulans* and one strain of *D. mauritiana*. The amount of variable sites in the polymorphism and divergence analysis are also shown.

Species (Sample Size)	Total Amount of Analyzable Nucleotides	Total Variable Sites
Drosophila simulans (8)	1173 bp	45 (40 singletons and 5 between two strains)
Drosophila mauritiana (1)	1201 bp	N/A
D. simulans and *D. mauritiana* (9)	1119 bp	84 (77 singletons and 7 between two strains/ species)

3.2 Within Species Polymorphism

Polymorphism analysis could only be done for *D. simulans* as I only had one *D. mauritiana* sequence. Eight strains of *D. simulans* were analyzed for polymorphisms within the species. Of the 1407 sites, there were 1,173 sites that had no gaps or missing data in any strain. There were 1,128 sites that were found to be invariable and 45 sites that were found to be variable. Within the variable sites, 40 were found to be singleton variable sites and 5 were shared between two strains. There were not any nucleotide sites had more than two variants. A sense of the possible role or lack of it, played by selection can be gained by comparing measures of the number of pairwise differences between sequences (π) and the proportion of segregating sites (θ). Pairwise differences are nucleotides in the aligned segments in which sequences differ. The proportion of segregating sites, is the proportion of sites that are polymorphic. The nucleotide diversity statistic, π is calculated by dividing the number of pairwise differences by the product of number of ways they can be paired and the number of

sites to compare. The value of π was determined to be 0.010, while θ was 0.015. The θ value is slightly higher than the π value, indicating an excess of low frequency single nucleotide polymorphisms (SNPs) (Krutovsky and Neale, 2005). Insertions and deletions were excluded from all estimates. These probabilities were based on all segregating sites, that is, sites which were polymorphic (Watterson, 1975).

The difference between π and θ forms the basis of Tajima's D test. Tajima's D test is a test used to detect deviations from neutrality from the distribution of polymorphisms. It does this by measuring the difference between the number of pairwise differences between sequences (π) and the proportion of segregating sites (θ) and determines if the result is in fact significant, otherwise the lack of significance supports neutrality. Tajima's D test value was determined to be -1.61 and not statistically different from zero ($0.10 > P > 0.05$). Fu and Li's D* test is based on the differences between the number of singletons (mutations appearing only once among the sequences) and the total number of mutations. Fu and Li's F* test is based on the differences between the total number of singletons and the average number of nucleotide differences between pairs of sequences (Fu and Li 1993), (Rozas et al. 2003). Fu and Li's D* test statistic was equal to -1.67 and Fu and Li's F* test statistic was -1.85. Both of these values were termed not significant as $P > 0.10$. A summary of these statistical results is shown in Table 9.

Table 9: Summary of the results of statistical tests based on polymorphism within *D. simulans*. π indicates the average number of pairwise differences, θ indicates the number of segregating sites, D indicates Tajima's D test, F* indicates Fu and Li's F test, and D* indicates Fu and Li's D test.

Gene	π	θ	D	F*	D*
CG14891	0.010	0.015	-1.61	-1.67	-1.85

3.3 Between Species Divergence

Eight strains of *D. simulans* were analyzed against one strain of *D. mauritiana* for divergent sites between the two closely related species. There were 84 variable sites, of which 41 consisted of fixed differences between the two species. Fixed differences refers to nucleotide differences that were only seen between species, and not within the *D. simulans* strains. The average number of nucleotide differences between the species was 49.0. The average number of nucleotide substitutions per site between species was 0.04379. The value of these numbers indicate that divergence at this gene appears to be rare between *D. simulans* and *D.*mauritiana, thus alterations are most likely either neutral or deleterious (Wayne and Simonsen, 1998).

The ratio of non-synonymous to synonymous polymorphism within *D. simulans* (Pi_a/Pi_s) and the ratio of non-synonymous to synonymous divergence between *D. simulans* and *D. mauritiana* (K_a/K_s) was graphed using a sliding window method (Figure 2). The graph indicates a small area near the 526 bp region which is very high in non-synonymous polymorphisms compared to the rest of the gene.

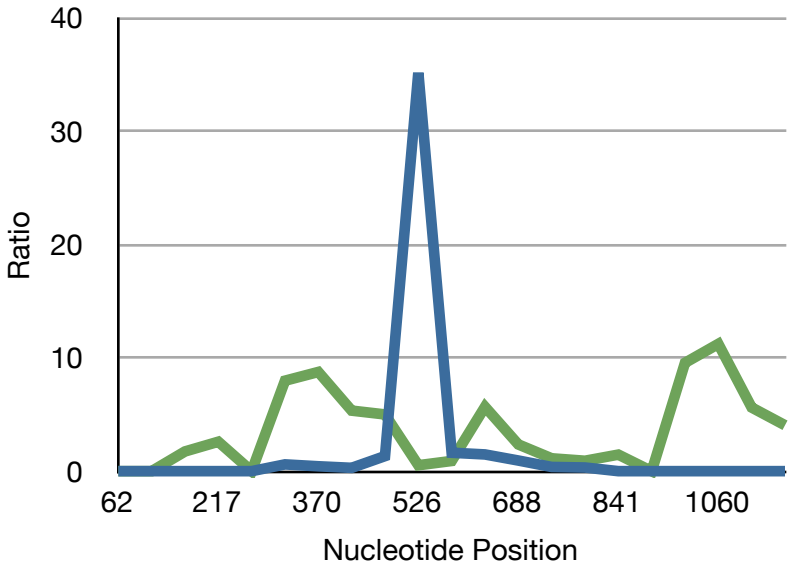

Figure 2: Sliding window analysis of the ratio of non-synonymous to synonymous polymorphism (Pi_a/Pi_s) within *D. simulans* (blue) and ratio of non-synonymous to synonymous divergence (K_a/K_s) between *D. simulans* and *D. mauritiana* (green). A window length of 100 sites and a step size of 50 sites was used. Nucleotide positions with alignment gaps were not counted in the window length and slide.

The genetic code is degenerate; this results in many amino acids being coded by multiple codons. It has been shown empirically that the use of optimal codons can increase the rate of translation (Kurland, 1987; Dey, 2004) and increase the fitness of an organism (Akashi, 1994). Non-synonymous mutations are point mutations that result in the coding of a different amino acid. Synonymous mutations are point mutations that result in the coding of the same amino acid. The neutral theory of molecular evolution assumes that synonymous mutations do not affect fitness (King and Jukes, 1969). Codon bias indicates differences in the frequency of occurrence of synonymous codons in DNA. Codon bias must be also accounted for when estimating d_N/d_S (ratio of non-synonymous to synonymous changes) as selection may favour synonymous codons;

thus a strong codon bias would yield a low d_S value. Codon bias was measured using the effective number of codons method (ENC). Synonymous codons are measured on a scale of 20-61 when compared for their codon bias. A value of 61 indicates no codon bias as there are 61 possible nucleotide combinations for amino acids. A ENC value of 60.167 was obtained, this value is extremely close to the value of 61, thus, no evidence of codon bias was found in either *D. simulans* or *D. mauritiana.*

The McDonald and Kreitman test is designed to test deviations from neutral evolution. This is accomplished by comparison of the number of amino acid replacement substitutions to synonymous substitutions in the coding region of a gene between species and comparing it to the numbers within species. If the observed substitutions are neutral, the non-synonymous to synonymous ratio of fixed differences between species should be equal to the ratio of non-synonymous to synonymous fixed differences within species. The test is well-suited for small sample sizes, even when the frequency of observations in one or more of the cells is less than five (Daya, 2002). The test assumes that synonymous substitutions are neutral, which in our case is a well supported assumption given the lack of codon bias.

The McDonald and Kreitman test was found to be nonsignificant (Fisher's exact test: P= 0.16) and thus the null hypothesis of neutrality was not rejected. When selection is weak, a longer time period is necessary for favourable mutations to become fixed within a species. The direction and degree of departure from the neutral prediction can be quantified with the neutrality index. It is interesting to note that the neutrality index was lower than one (NI= 0.40) indicating a larger proportion of non-synonymous substitutions between species than within species.

Table 10: Results of the McDonald and Kreitman Test for the coding region of the candidate gene CG14891 in *D. simulans* (n=8) and *D. mauritiana* (n=1). Fixed differences between species refer to both *Drosophila* species, whereas polymorphic sites refers to only *D. simulans*.

Substitution Type	Fixed Differences Between Species	Polymorphic Sites
Synonymous Substitutions	5	11
Non-synonymous Substitutions	35	31

Despite the lack of significance for both tests of polymorphism within and between species it is worth noting that the tests suggests low frequency of polymorphisms within species. This could be linked to some purging effect by purifying selection; however a trend towards a larger proportion of non-synonymous substitutions between species than within species was found, suggesting possible weak positive selection, that is, genetic drift between species. In fact, it is interesting to note that the ratio of non-synonymous to synonymous polymorphism was slightly lower than one ($Pi_a/Pi_s = 0.81$) in line with expectations from weak purifying selection or neutrality, whereas the same ratio for comparisons between species was higher than one ($K_a/K_s = 2.1$) in agreement with positive or adaptive selection.

Table 11: The ratio of non-synonymous to synonymous polymorphisms calculated for *D. simulans* within species and the ratio of non-synonymous to synonymous nucleotide changes regarding between species divergence for *D. simulans* and *D. mauritiana.* A value of 1 indicates neutrality.

Test	Ratio of Non-synonymous to Synonymous Nucleotide Changes
Within Species (*D. simulans)*	0.81
Between Species (*D. simulans* and *D. mauritiana*)	2.1

4.0 Discussion

It was determined that polymorphisms within species were found to be rare, and a trend of adaptive divergence between species was found within the candidate gene CG14891. However only four-fifths of the gene CG14891 was successfully sequenced. Factors that contributed to the inability to obtain sequence data for the entire gene included incomplete sequencing results as well as the inability to differentiate between nucleotide bases for single position. There are many possible explanations for the inability to obtain complete sequence data. One possible explanation is that the incomplete sequence runs were a result of poor template concentration estimation. Although a fairly large range of mass of DNA is known to work efficiently in the sequencer, estimation of the specific mass was done by eye and simply by the comparison of intensity to the ladder used. A sample that contains too little template is known to lead to poor signal strength and thus poor base calling and shorter sequences; whereas too much DNA template results in shorter read lengths. This is due to a greater number of fragments being preferentially loaded into the capillary during sequencing (Calvallin, 2010). A second possible explanation is that since it is known that there is a generally lower reading accuracy in homopolar stretches of identical bases (Ansorge, 2009). Long homopolar A/T regions in cloned DNA can result in problems in DNA sequencing reactions due to polymerase slippage in the repeating region. This is thought to occur because the two strands do not stay paired correctly during polymerization; this generates repeat fragments of varying lengths that are still able to retain the correct sequence after the repeat (Calvallin, 2010). However, this

does not seem to be the case in any of the sequences used because there were no long A/T regions in the sequence data.

Biological replicates were used for sequencing as they protect against both rare mutations and possible errors in the extraction procedure. However, sequencing using forward and reverse primers from two biological replicates leads to a small number of cases in which there was no agreement in base calls for a particular nucleotide position. One possible explanation is that these positions were areas of heterozygosity within the strain. These sites were termed 'N' and removed from the analysis. The placement of some primers especially at the end of the coding region were inefficient at obtaining sequencing results for the extreme end region of the gene. Specifically, the placement of a forward primer instead of one of the two reverse primers in the final third of the gene may have aided the process and allowed a greater portion of sequence data to be available.

In a large prior study in which 11,466 genes of seven strains of *D. simulans* were sequenced, approximately 98% of the coding sequence, that is a total of 393,951,345 base pairs were successfully sequenced. From this large analysis, 2,965,987 polymorphic nucleotides were determined (Begun et al. 2007). This corresponds to approximately 7.5 polymorphic nucleotides for every 1000 total nucleotides. My study focused on a single gene and used a slightly higher number of strains of *D. simulans*. A total of 1,173 base pairs were successfully sequenced and found to have 1,128 invariable sites. This corresponds to approximately 38.4 polymorphic nucleotides for every 1000 total nucleotides. Thus a 5 fold increase in polymorphic nucleotides were found in the candidate gene CG14891, than in the average gene in the genome of *D.*

simulans. As well, this same study determined that of the 2,965,987 polymorphic nucleotides, 43,878 coded for a change in the amino acid (Begun et al. 2007). Of the 45 polymorphic sites that were found within the coding region of this CG14891, 31 coded for a change in amino acid, where as 14 did not result in an amino acid change. Thus, in the analysis of the entire genome of *D. simulans*, only 1.48% of the nucleotide changes resulted in a change in amino acid. In contrast, on the candidate gene CG14891, a large proportion, that is 68.9%, of the nucleotide changes coded for a different amino acid.

A high number of polymorphic changes as well as non-synonymous nucleotide changes were not unexpected because CG14891 was not a randomly chosen gene, rather a candidate gene for CSP. Several other *Drosophila* genes that have in the past been targets for molecular population genetic analysis were targeted based on their presumed association with a phenotype theorized to have a history of adaptive evolution (Begun et al. 2007). These genes, including CG14891, will likely produce differing data compared to the entire *Drosophila* genome in regards to the proportion of proteins diverging under adaptive evolution (Begun et al. 2007). As well, since reduced crossing over throughout the genus *Drosophila* is known to reduce levels of polymorphism while still maintaining normal levels of divergence between species (Begun et al. 2007); these results may indicate that CG14891 experiences high rates of crossing over. As well, genes that encode for sex and reproduction related proteins are known to be subject to a large array of selective forces, such as sperm competition, sexual conflict, and conspecific sperm precedence. These selective forces are of evolutionary consequence and should theoretically lead to the rapid evolution of these

genes (Civetta and Singh,1995; Haerty et al. 2007). Prior work within the *D. melanogaster* sub-group analyzed 2,505 sex and reproduction related genes and they did find that male sex and reproduction-related genes evolve quicker at the protein level than genes not involved in sex or reproduction (Haerty et al. 2007).

Another way to compare polymorphisms at this gene is to analyze the number of pairwise differences between sequences and the proportion of segregating sites. In prior research, a set of 14 genes, all located on the third chromosome of *Drosophila simulans* were sequenced in a number of strains and the pairwise differences between sequences and the proportion of segregating sites were calculated (Kern et al. 2002)**.** The values in Table 12, correspond to the nucleotide diversity values of this prior experiment, as well as from CG14891. The θ in CG14891 can be shown to be almost equivalent to the average proportion of segregating sites for the fifteen genes on the same chromosome. As well, π was determined to be only slightly higher than the average for the subset of third chromosome genes.

Table 12: Estimates of nucleotide heterozygosity for 15 genes located on the third chromosome of *D. simulans*. n, sample size; Chrm., chromosome arm; Cyt, cytological location; θ and π, estimates of nucleotide heterozygosity based on number of segregating sites and number of pairwise differences. Note, all results from loci except CG14891 attained from Kern et al. 2002.

Locus (n)	Chrm.	Cyt.	θ	π
CG14891(8)	3R	89B	0.015	0.01
AP-50 (8)	3R	94A	0.0160	0.0161
boss (5)	3R	96F	0.0116	0.0118
Cen190(7)	3R	88E	0.0073	0.0091
eld(7)	3R	90B	0.0082	0.0107
Est-6 (4)	3R	68F	0.0178	0.0182
Hsc70-1 (7)	3R	70C	0.0032	0.0031
hyd (8)	3R	92E	0.0032	0.0031
mira (6)	3R	86A	0.0044	0.0056
Osbp (8)	3R	96B	0.0112	0.0113
Rel (7)	3Q	85C	0.0077	0.0069
Rh3(5)	3R	85F	0.0127	0.0127
ry(18)	3R	90A	0.0186	0.0192
T-cpl (8)	3R	94A	0.0029	0.0028
Tpi (9)	3R	99D	0.0087	0.0068
Average			**0.0099**	**0.0098**

Based on the number of polymorphisms determined within *D. simulans* for the eight strains analyzed, the result indicates that polymorphisms although high compared to the entire genome appear to be similar to polymorphism data for other third chromosome genes. The fact that the estimate for segregating sites is higher than the

estimate for pairwise differences reflects the fact that most nucleotide changes are rare. Such a pattern is expected for neutral variables that are simply a consequence of recent mutations, with most of them lost by random genetic drift and possibly purifying selection removing deleterious mutations.

A drawback of Tajima's D test is that the number of segregating sites in a sample is only an asymptotic sufficient statistic for θ under the infinite sites model and the efficiency of using segregating sites to estimate θ can be extremely low for finite samples (Fu, 1995). Fu and Li's tests have been shown to be slightly less conservative than Tajima's D test, however the predefined θ interval provides the opportunity for the test to falsely reject neutrality when θ is not in this interval. As selection is a diverse and complex phenomenon, there is not one test that can be considered the most powerful (Wayne and Simonsen, 1998). Fu and Li's tests may also be advantageous in cases where there is evidence of codon bias (Wayne and Simonsen, 1998). However, this was not the case here as no codon bias was found.

One drawback of neutrality tests is their assumption of an infinite-sites model with no recombination. The evolution of neutral sites is not altered by weak purifying selection even if no recombination is detected between them; as well, variation at these sites is reduced due to background selection on deleterious mutations (de Brito et al. 2002). However, data analysis assuming free recombination (data not shown) did not alter the results of the tests. Both Fu and Li's tests as well as Tajima's test were found to be non-significant and therefore the results suggest neutrality within species. However the tests are known to be highly conservative and so it is worth paying attention to the

direction of the test. The negative sign of the estimates suggest a possible role for weak purifying selection removing deleterious mutations.

A highly variable region of the candidate gene in terms of polymorphisms is located around the 526 bp region. This is interesting as it suggests two possible explanations; an area with a high mutation rate and/or low functional constraint, or a region subjected to balancing selection. If the area has a high mutation rate and/or low functional constraint, the amino acid polymorphisms should have similar chemical properties and should be able to function and interact in similar ways suggesting neutrality. However, the polymorphisms involved have radical (amino acids that have remarkably different chemical properties) rather than conservative amino acid changes (Appendix 3). Therefore, this may indicate an area of the gene with large variation driven by adaptations to sperm competitive ability within species.

My estimate of codon bias agrees with prior studies on *Drosophila melanogaster*, which argue that the strength of selection for any synonymous codon substitution would be negligible in this species as the difference in fitness which would arise out of preferential usage of optimal codons was likely to be extremely low (McVean and Viera, 2001). The G + C content for the entire coding region was 0.456; this indicates an equal usage of all nucleotides. This is in agreement with prior results suggesting that mutational bias does not play a role in codon usage bias in *D. melanogaster* (Powell and Moriyama, 1997). The lack of mutational bias and codon bias makes it possible to rely on high d_N/d_S estimates or an excess of non-synonymous replacements in the McDonald and Kreitman test, as indicative of adaptive selection.

The protein sequence of *D. simulans* for CG14891 was analyzed for the presence of specific motifs. There were no clear motifs present, however the possibility of many motifs such as cAMP- and cGMP-dependent protein kinase phosphorylation site, Casein kinase II phosphorylation site, N-myristoylation site, and Protein kinase C phosphorylation site, were present. Interestingly, a N-myristoylation site and several putative phosphorylation sites are known to be found on the gene tescalcin found in mice and is expressed in the early stages of testes differentiation (Perera et al. 2001). In mice, this gene functions in testis development, signal transduction, control of cell cycle and differentiation (Perera et al. 2001). Thus in *Drosophila,* it is possible that similar effects may occur due to CG14891 and that alterations in this gene may in turn affect the competitive ability of the sperm produced by the testes. The similarities between the *Drosophila* and mouse protein may thus indicate they are involved in similar mechanisms. However this is a fairly large leap as there were no motifs that can be confirmed in the candidate gene.

Future work should include sequencing the final one fifth of CG14891 in all strains such that the entire gene can be used in analysis. As well, an increase in strains, particularly in *D. mauritiana* will be needed to justify the results. Future work should also include sequencing all other candidate genes found for conspecific sperm precedence, particularly those in the 89B region of the mapped chromosome.

5.0 Conclusions

- Overall, polymorphisms within species were found to be rare, and a trend of adaptive divergence between species was found.

- The statistical tests however, showed that these results were non-significant and thus neutrality can not be rejected.

- Therefore, even though the trend in regards to polymorphism and divergence is noteworthy, the nonsignificant results of the statistical results hinder the ability to fully validate this gene.

6.0 References

Adams, E. M., and Wolfner, M. F., 2007. Seminal proteins but not sperm induce morphological changes in the *Drosophila melanogaster* female reproductive tract during sperm storage. Journal of Insect Physiology **53**: 319-331.

Akashi, H., 1994. Synonymous Codon Usage in *Drosophila melanogaster*: Natural Selection and Translational Accuracy. Genetics **136**: 927–935.

Ansorge, W. J., 2009. Next-generation DNA sequencing techniques. New Biotechnology **25**: 195-203.

Begun, D. J., Holloway, A. K., Stevens, K., Hiller, L. W., Poh, Y. P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., and Langley, C. H., 2007. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in *Drosophila simulans.* PloS Biol. **5**: 2534-2559.

Birkhead, T. R., 1998. Cryptic Female Choice: Criteria for Establishing Female Sperm Choice. Evolution **52**: 1212-1218.

Britch, S. C., Swartout, E. J., Hampton, D. D., Draney, M. L., Chu, J., Marshall, and J. L., Howard, D., 2007. Genetic Architecture of Conspecific Sperm Precedence in *Allonemobius fasciatus and A. socius.* Genetics **176**: 1209-1222.

Calvallin, K. 2010. Iowa State University of science and technology, DNA facility, Ames, Iowa http://www.dna.iastate.edu/index.html

Chang, A.S., 2004. Conspecific Sperm Precedence in Sister Species of *Drosophila* with Overlapping Ranges Evolution **58**: 781-789.

Civetta, A., and Singh, R. S., 1995. High divergence of reproductive tract proteins and their association with post-zygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. J. Mol. Evol. **41**: 1085-1095.

Civetta, A., 1999. Direct visualization of sperm competition and sperm storage in *Drosophila.* Current Biology **9**: 841-844.

Civetta, A. and Singh, R. S. 1999. Broad-sense sexual selection, sex gene pool evolution, and speciation. Genome **42**: 1033-1041.

Civetta, A., Waldrip-Dail, H. M., and Clark, A. G., 2002. An introgression approach to mapping differences in mating success and sperm competitive ability in *Drosophila simulans* and *D. sechellia*. Genet. Res. Camb., **79**: 65-74.

Civetta, A., Rosing, K. R., and Fisher, J. H., 2008. Differences in sperm competition and sperm competition avoidance in D*rosophila melanogaster*. Animal Behavior **75**: 1739-1746.

David, J. R., Lemeunier, F., Tsacas, L., and Yassin, A., 2007. The Historical Discovery of the Nine Species in the *Drosophila melanogaster* Species Subgroup. Genetics **177**: 1969-1973.

Daya, S., 2002. Understanding Statistics: Fisher exact test. Evidence-based Obstetrics and Gynecology **4**: 3-4.

de Brito, R. A., Manfrin, M. H., and Sene, F. M., 2002. Mitochondrial DNA phylogeography of Brazilian populations of *Drosophila buzzatii*. Genetics and Molecular Biology **25**: 161-171.

Dey, S., 2004. Benefits of being biased! Journal of Genetics **83**: 113-115.

Drosophila 12 Genomes Consortium, 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature **450:** 203-218.

Fu, Y. X., and W. H. Li., 1993. Statistical tests of neutrality of mutations. Genetics **133**: 693–709.

Fu, Y.X., 1995. Statistical Properties of Segregating Sites, Theoretical Population Biology **48**: 172-197.

Gilbert, W. and Maxam, A., 1973. The Nucleotide Sequence of the *lac* Operator. Proc. Nat. Acad. Sci. **70:** 3581-3584.

Halliburton, R., 2004. Introduction to Population Genetics. Pearson Education, Inc. New Jersey, USA.

Harshman, L. G. and Clark, A. G., 1998. Inference of sperm competition from broods of field-caught *Drosophila*. Evolution **52**: 1334–1341.

Haerty, W., Jagadeeshan, S., Kulathinal, R. J., Wong, A., Ram, K., R., Sirot, L., K., Levesque, L., Artieri, C. G., Wolfner, M. F., Civetta, A., and Singh, R. S., 2007. Evolution in the Fast Lane: Rapidly Evolving Sex-Related Genes in *Drosophila.* Genetics **177**: 1321-1335.

Holman, L., 2009. *Drosophila melanogaster* seminal fluid can protect the sperm of other males. Functional Ecology **23**: 180-186.

Kern, A. D., Jones, C., D., and Begun, D. J., 2002. Genomic Effects of Nucleotide Substitiutions in *Drosophila simulans* Genetics **162**: 1753-1761.

Kilman, R. M, Andolfatto, P., Coyne, J. A., Depaulis, F., Kreitman, M., Berry, A., McCarter, J., Wakeley, J., and Hey, J., 2000. The Population Genetics of the Origin and Divergence of the *Drosophila simulans* Complex Species. Genetics **156**: 1913-1931.

Kimura, M., and Crow, J. F., 1964. The number of alleles that can be maintained in a finite population. Genetics **49**: 725-738.

King J. L., and Jukes, T. H., 1969. Non-Darwinian Evolution. Science **164**: 788-798.

Kreitman, M., 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature **304**: 412-417.

Krutovsky, K. V., and Neale, D. B., 2005. Nucleotide Diversity and Linkage Disequilibrium in Cold-Hardiness-and Wood Quality-Related Candidate Genes in Douglas Fir. Genetics **171**: 2029-2041.

Kurland, C.G., 1987. Strategies for efficiency and accuracy in gene expression. 1. The major codon preference: a growth optimization strategy. Trends Biochem. Sci. **12**: 126-128.

Levesque, L., Brouwers, B., Sundararajan, V., and Civetta, A., 2010. Third chromosome candidate genes for conspecific sperm precedence between *D.simulans* and *D.mauritiana*. BMC Genetics **11**: 21- 57.

McDonald, J. H., and Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature **351**: 652–654.

McVean, G.A.T. and Viera, J., 2001. Inferring Parameters of Mutation, Selection and Demography From Patterns of Synonymous Site Evolution in *Drosophila*. Genetics **157**: 245–257.

Ohnishi, S., Kawanishsi, M., and Watanabe, T. K., 1983. Biochemical phylogenies of *Drosophila:* protein differences detected by two-dimensional electrophoresis. Genetica **61**: 55-63.

Perera, E. M., Martin, H., Seeherunvong, T., Kos, L., Hughes, I. A., Hawkins, J. R., and Berkovitz, G. D., 2001. Tescalcin, a novel gene encoding a Putative EF-Hand Ca^{2+} - Binding Protein, Col9a3, and Renin Are Expressed in the Mouse Testis during the Early Stages of Gonadal Differentiation. Endocrinology **142**: 453-463.

Pitnick, S., 1996. Investment in Testes and the Cost of Making Long Sperm in *Drosophila* **148**: 57-80.

Powell, J.R and Moriyama, E.N., 1997. Evolution of codon bias in *Drosophila*, Proc. Natl. Acad. Sci. **94**: 7784-7790.

Price, C.S.C., 1997. Conspecific Sperm Precedence in *Drosophila*. Nature **388**: 663-666.

Price, C.S.C., Dyer, K.A., and Coyne, J. A., 1999. Sperm competition between *Drosophila* males involves both displacement and incapacitation. Nature **400**: 449-452.

Price, C.S.C., Kim, C.H., Posluszny, J., and Coyne, J. A., 2000. Mechanisms of Conspecific Sperm Precedence in *Drosophila*. Evolution **54**: 2028-2037.

Price, C. S. C., Kim, C. H., Gronlund, C. J., Coyne, J. A., 2001. Cryptic Reproductive Isolation in the *Drosophila simulans* Species Complex. Evolution **55**: 81-92.

Ramm, S. A., Parker, G. A., and Stockley, P., 2005. Sperm competition and the evolution of male reproductive anatomy in rodents. Proc. R. Soc. B. **272**: 949-955.

Rozas, J., J. C. Sánchez-DelBarrio, X. Messeguer, and R. Rozas., 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19**: 2496–2497.

Sanger, F., Nicklen, S., and Coulson, A. R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. **74**: 5463-5467.

Simmons, L. W., 2001. Sperm Competition and its Evolutionary Consequences in the Insects. Princeton, New Jersey: Princeton University Press.

Singh, S. R., Singh, B. N., Hoenigsberg, H. F., 2002. Female re-mating, sperm competition and sexual selection in *Drosophila.* Genet. Mol. Res. **1**: 178-215.

Snook, R. R., and Markow, T. A., 2002. Efficiency of gamete usage in nature: sperm storage, fertilization and polyspermy. Proc. R. Soc. Lond. B **269**: 467-473.

Solignac, M., and Monnerot, M., 1986. Race Formation, Speciation, and Introgression with *Drosophila simulans, D. mauritiana*, and *D. sechellia* inferred from mitochondrial DNA analysis. Evolution **40**: 531-539.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

Tao, Y., Chen, S., Hartl, D. L., and Laurie, C. C., 2003. Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritiana.* I. Differential accumulation of hybrid male sterility effects on the X and autosomes. Genetics **164**: 383-397.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G., 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **24**: 4876-4882.

Watterson, G. A., 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**: 256–276.

Wayne, M. L., Simonsen, K. L., 1998. Statistical tests of neutrality in the age of weak selection. Trends in Ecology and Evolution **13**: 236-240.

Yang, Z., 1997. PAML; A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**: 555-556.

7.0 Appendicies

Appendix 1:

Forward and reverse primer sequences used during polymerase chain reaction and sequencing reaction. Primers are aligned on the coding region of *Drosophila simulans*. Note that the reverse PCR and sequencing primer was from a noncoding portion of the sequence. Lower case letters refer to the regions outside the coding region of CG14891.

TATGTTCCGCTTGAAGGGAAATTTTGGAAATTTTTACTTTAAAAATGGTTTGGTGTAC
ACCAAAGATCGC**AAGGTTGTGCGACTGGTGA**[1]CGATCTCTGCGAACTGGCACTTTT
CGAAGAGCCAACTCTGGAAGCACTTCTCGAGTTTTGGAACTGTGGAGGATCTCCAA
TGGGAAAAGGATAAGAGAGTGGGATCGGTTCTTTTTCAAGAGGCTTCCCAAGCGG
CAAAGGTTTTGGTGTTGACTAAACACCATTTGTATGGCCATGTTCTTTATTTGCAGCC
CAGCACCTCCAGGCGCGAACCGCCGGTGAAGGAATCAGAAACTATTTCTGCCTAC
GATATACCCGTTGTCGATGACTTTTGGTATAAAGTGCTCGAATAT**CTTCCACTAAATG**
CCCGTCT[2]CAACTTTGCCGCCAGTTGTAAAAGATTCAAGACGATCTACG**AATTGGAG**
TCGCGTCGTAAC[3]AATCGTGTTCTTAATATGAAGGATGTTTGCACACTGGACGACTTT
GGCATTAAAATATT**GATGCGGCTATCAGGAAAAC**[4]ACATTCATTGTGTAAAAGGTGG
CCCGCTTCATTGGACGCTTATGT**TGGAGTTCGTGCAGCTATTG**[5]GGTGTAAGCTGTC
CAAATCTAGCAGAGCTAAGTTTCTACAAAATTTCAGTCAGCCTAGACCACATGACTC
ACCTGTTCGATGGTGCCAATGGCTTGAATAATATCACCACCATATCCTTGAGGTGTT
GTGACTTGGCAGATCCTCAAATTTACTGCTTGCAGATGCTATCTAAACTAAAGAGTCT
GGACATCGCACAGAACCATTTCATTAGGGGCGAAAGTTTAAACTCTCTGCCAATTTC
CTTGGAGATTTTAAATGTTTCAAAATGCGACAGACTGCGGCCCAAGAATCTTATCAAT
CTTGCGTCCCTGACGCATCTCCGCGAACTGCGCTGCTCTGGCATTTCCAAGCTTAC
GAAAAATGAGCTGTTCAAACGGTTCGCACATTACT**GTCCAATGCTCGA*GGTTCTC*[6]G**
AGGTTACCGACA[7]TTATGAAGAAGATACAGCTGGGCGGTCTGTCTCGTCTCCACAC
CTTGGTCATTCAGTCTTCCGAAGGGTCTGGCGACCATATGAATAACTTGATGCTTTC
GTCGATCGCGGAATCGTATTCGCTGCGCCGTCTGGAGATTATAGATTCTTTTGAACG
TTTTTTCACTATTTCCTTCGATCTGAGTATTTTATCCCCGCTTAAAGAACTGCGGACC
CTAATATTACATAATCTGAACTTTACACCGGAACACCTAATGGGATTGCAAAAACTCC
CTGCCTTGGAG**TTTCTGGACCTGAGTGGCTC**[8]GCCCGATCTATCCAATGAGGACGT
TGCAAAGTTGACGAAACCGCTGGGCAGACTGCGCCGACTAACGGTTGAGCGTTGT
CCTTTTATCTCACGACAATTAACCGAGATCCTGAAAGGGAATCCTAAGCTTCAAGTA
GTCTTTTAGggtcagctcaacaacaatgtttttatatcatt**cctaagaagtgcagactttgca**[9]tttgtatcaatttgtatc
aactaactttatattcgacccgctgagcttatggaataaattgtccattttgtaatattgttaactccaccgccaaaatataataa
atacatacagata

LEGEND:

[1] (F_yel) PCR and sequencing forward primer (*AAGGTTGTGCGACTGGTGA*)

[2] (I_417) Sequencing and reverse primer (*AGACGGGAATTTAGTGGAAG*) Reverse Complement:(*CTTCCACTAAATGCCCGTCT*)

[3] (I_621) Sequencing forward primer *(AATTGGAGTCGCGTCGTAAC)*

[4] (Prim_1) Sequencing forward primer *(GATGCGGCTATCAGGAAAAC)*

[5] (Prim_2) Sequencing reverse primer (*CAATAGCTGCACGAACTCCG)*
Reverse Complement: *(**C**GGAGTTCGTGCAGCTATTG)*

[6] (Prim_3) Sequencing reverse primer *(GAGAACCTCGAGCATTGGAC)* Reverse Complement: *(GTCCAATGCTCGAGGTTCTC)*

[7] (Prim_4) Sequencing reverse primer (*TGTCGGTAACCTCGAGAACC*)
Reverse Complement: *(GGTTCTCGAGGTTACCGACA)*

[8] (Fe_R) Sequencing reverse primer *(GAGCCACTCAGGTCCAGAAA)*
Reverse Complement: *(TTTCTGGACCTGAGTGGCTC)*

[9] (R_grn) PCR and sequencing reverse primer *(TGCAAAGTCTGCACTTCTTAGG)* Reverse Complement: *(CCTAAGAAGTGCAGACTTTGCA)* Note: this comes after coding region.

Appendix 2:

A list of non-synonymous codon changes in the analyzable region of the candidate gene CG14891 between eight strains of *D. simulans.*

Site	Consensus (charge/ polarity)	Rare (charge/ polarity)	Strain	Site	Consensus (charge/ polarity)	Rare (charge/ polarity)	Strain
13	Val (neutral/ non-polar)	Met (neutral/ non-polar)	3	559	Ala (neutral/ non-polar)	Thr (neutral/ polar)	6
47	Lys (positive/ polar)	Arg (positive/ polar)	3	566	Leu (neutral/ non-polar)	Gln (neutral/ polar)	15
88	Glu (negative/ polar)	Gln (neutral/ polar)	210	595	Asp (negative/ polar)	His (neutral/ polar)	15
109	Asp (negative/ polar)	Asn (neutral/ polar)	106, 03	657	Leu (neutral/ non-polar)	Phe (neutral/ non-polar)	57
130	Leu (neutral/ non-polar)	Phe (neutral/ non-polar)	63	676	Asp (negative/ polar)	His (neutral/ polar)	6
226	Arg (positive/ polar)	Gly (neutral/ non-polar)	63	684	Gln (neutral/ polar)	His (neutral/ polar)	3
324	Asn (neutral/ polar)	Lys (positive/ polar)	6	688	Tyr (neutral/ polar)	His (neutral/ polar)	3
362	Lys (positive/ polar)	Met (neutral/ non-polar)	210	691	Cys (neutral/ non-polar)	Gly /Ala (neutral/ non-polar)	106
400	Val (neutral/ non-polar)	Leu (neutral/ non-polar)	3	692	Cys (neutral/ non-polar)	Ser /Ala (neutral/ polar /non-polar)	106
458	Arg (positive/ polar)	Leu (neutral/ non-polar)	106	699	Gln (neutral/ polar)	His (neutral/ polar)	6,106
508	Thr (neutral/ polar)	Pro (neutral/ non-polar)	106	710	Lys (positive/ polar)	Ile (neutral/ non-polar)	3
516	Met (neutral/ non-polar)	Ile (neutral/ non-polar)	6	736	Asn (neutral/ polar)	Asp (negative/ polar)	63

Site	Consensus (charge/ polarity)	Rare (charge/ polarity)	Strain	Site	Consensus (charge/ polarity)	Rare (charge/ polarity)	Strain
518	Leu (neutral/ non-polar)	Ser (neutral/ polar)	15	744	Phe (neutral/ non-polar)	Leu (neutral/ non-polar)	199
519	Leu (neutral/ non-polar)	Phe (neutral/ non-polar)	106	755	Glu (negative/ polar)	Gly (neutral/ non-polar)	63
541	Val (neutral/ non-polar)	Leu (neutral/ non-polar)	57	824	Lys (positive/ polar)	Arg (positive/ polar)	6
545	Ser (neutral/ polar)	Ile (neutral/ non-polar)	15				

Appendix 3:

A list of non-synonymous codon changes in the analyzable region of the candidate gene CG14891 between the eight strains of *D. simulans* and the single strain of *D.* mauritiana.

Site	*D. simulans* (charge/ polarity)	*D. mauritiana* (charge/ polarity)	Site	*D. simulans* (charge/ polarity)	*D. mauritiana* (charge/ polarity)
103	Glu (negative/ polar)	Lys (positive/ polar)	679	Pro (neutral/non-polar)	Thr (neutral/ polar)
158	Lys (positive/ polar)	Arg (positive/ polar)	749	Arg (positive/ polar)	Lys (positive/ polar)
161	Val (neutral/non-polar)	Gly (neutral/ non-polar)	765	Asn (neutral/ polar)	Lys (positive/ polar)
172	Thr (neutral/ polar)	Ala (neutral/ non-polar)	813	Arg (positive/ polar)	Ser (neutral/ polar)
181	His (neutral/ polar)	Tyr (neutral/ polar)	818	Arg (positive/ polar)	Leu (neutral/ non-polar)
216	Ser (neutral/ polar)	Arg (positive/ polar)	826	Asn (neutral/ polar)	His (neutral/ polar)
223	Arg (positive/ polar)	Gly (neutral/ non-polar)	850	Thr (neutral/ polar)	Ser (neutral/ polar)
227	Arg (positive/ polar)	His (neutral/ polar)	908	Phe (neutral/ non-polar)	Tyr (neutral/ polar)
251	Glu (negative/ polar)	Gly (neutral/ non-polar)	928	Cys (neutral/ non-polar)	Ser (neutral/ polar)
256	Ile (neutral/ non-polar)	Leu (neutral/ non-polar)	962	Ile (neutral/ non-polar)	Thr (neutral/ polar)
289	Phe (neutral/ non-polar)	Val (neutral/non-polar)	1031	Ser (neutral/ polar)	Phe (neutral/ non-polar)
325	Ala (neutral/ non-polar)	Ser (neutral/ polar)	1054	Met (neutral/ non-polar)	Leu (neutral/ non-polar)

Site	*D. simulans* (charge/ polarity)	*D. mauritiana* (charge/ polarity)	Site	*D. simulans* (charge/ polarity)	*D. mauritiana* (charge/ polarity)
340	Ala (neutral/ non-polar)	Ser (neutral/ polar)	1060	Ser (neutral/ polar)	Thr (neutral/ polar)
386	Arg (positive/ polar)	His (neutral/ polar)	1091	Arg (positive/ polar)	His (neutral/ polar)
415	Asp (negative/ polar)	Tyr (neutral/ polar)	1125	Phe (neutral/ non-polar)	Leu (neutral/ non-polar)
418	Val (neutral/non-polar)	Ile (neutral/ non-polar)	1126	Thr (neutral/ polar)	Ala (neutral/ non-polar)
505	Trp (neutral/ non-polar)	Arg (positive/ polar)	1160	Leu (neutral/ non-polar)	Pro (neutral/ non-polar)
508	Thr (neutral/ polar)	Ala (neutral/ non-polar)			