Summer 8-2011

Putative Protamine-like Proteins in 12-Sequenced Drosophila Species

Zain A. Alvi
Seton Hall University

Putative protamine-like proteins in 12 sequenced Drosophila species

Zain A. Alvi

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in Biology, Neuroscience Track from the Department of Biological Sciences of Seton Hall University

August 2011

APPROVED BY

CO-MENTOR

Dr. Angela V. Klaus

CO-MENTOR

Dr. Tin-Chun Chu

COMMITTEE MEMBER

Dr. Heping Zhou

Carolyn S Bentivegna

COMMITTEE MEMBER

Dr. Carolyn Bentivegna

DIRECTOR OF GRADUATE STUDIES

Dr. Allan Blake

CHAIRPERSON, DEPARTMENT OF BIOLOGICAL SCIENCES

Dr. Jane Ko

Acknowledgements

I would like to extend my deepest and sincere gratitude to the following people:

Dr. Angela Klaus, my mentor and thesis advisor, whose support, guidance, and oversight throughout the course of my research of my undergraduate and graduate years at Seton Hall University was invaluable. I am forever grateful and appreciative for the countless hours and effort that Dr. Klaus has spent throughout the years. Her constant encouragement and dedication to aid and motivate me in tackling our project even when we ran into some minor setbacks will have an eternal impact on me to become a better and a well versed researcher. In addition to my thesis, I would like thank Dr. Klaus for the insightful discussions about graduate school life as a whole. It truly was a great honor to have worked under guidance of such gifted, multi-versed, and talented scientist as Dr. Klaus. My deepest appreciation and thanks for affording me the opportunity to work alongside her.

Dr. Tin-Chun Chu, my mentor and thesis advisor, whose support and guidance throughout the course of my research as well as in class at Seton Hall University was invaluable. I am truly forever grateful for Dr. Chu investing a lot of time in improving my understanding of the many bioinformatics tools available and how to properly and efficiency use the molecular techniques on our project. In addition, I am sincerely appreciative of all the encouragement and continuous support throughout the project. Aside from my thesis, I would like to thank Dr. Chu for the wonderful discussion that we had about graduate school life as a whole. My deepest appreciation and thanks for affording me the opportunity to work alongside her.

Dr. Carolyn Bentivegna, chairperson of the Department of Biological Sciences at Seton Hall University, whom I owe many thanks for her insightful discussions about not only my research but also my undergraduate and graduate school life as whole. I also thank her for allowing me to serve as a teaching assistant for the Department of Biological Sciences and to perform research within its laboratories. My sincere and deepest gratitude to her for her continuous encouragement and guidance throughout my academic career at Seton Hall University and for sitting on my committee.

Dr. Heping Zhou, for sitting on my committee and for wonderful guidance concerning molecular techniques in the classroom and in the laboratory at Seton Hall University. I sincerely appreciate her wonderful help.

Dr. Allan Blake, whom I owe much thanks for his insightful discussions about not only my thesis work, but about graduate school life as a whole. My sincere and deepest gratitude to him for his continuous encouragement and guidance throughout my academic career at Seton Hall University.

Dr. Carroll Rawn, whom I owe many thanks for the privilege to do my first research project in the field of Bioinformatics and Computational Biology as undergraduate student at Seton Hall University. My first project opened a whole new world of scientific research for which Dr. Rawn has my deepest gratitude.

Dr. Jane Ko, whom I owe many thanks for her insightful discussions about not only thesis work, but also about graduate school as a whole. My sincere and deepest gratitude to her for the continuous encouragement to continue doing research in bioinformatics.

The Department of Biological Sciences Faculty and Staff, who not only provided valuable institution in the classroom, but kept me advancing and supportive in my academic pursuits over the past six years at Seton Hall University.

Dr. Valerie Schawaroch, whom I owe much thanks for taking time from her busy schedule to provide insight on the process of extracting DNA from and conducting PCR with a single Drosophila fly at – Baruch College (CUNY). I sincerely appreciate her wonderful help.

Dr. Sara Guariglia, whom I owe much thanks for spending many hours of her valuable time teaching me the use of the ultramicrotome for preparing transmission electron microscopy samples at College of Staten Island (CUNY). Also her insight about improving our fixation and dehydration protocols is appreciated. I am sincerely grateful her wonderful help.

My parents and friends, whose immense love, support, and continuous encouragement is truly been immeasurable.

Table of Contents

List of Figures

List of Tables

Abstract

The current study is aimed at analyzing putative protein sequences of the protamines of 12 Drosophila species based upon the reference sequences of two protamines-like proteins (Mst35Ba and Mst35Bb) found in *Drosophila melanogaster* sperm nuclei. Protamine-like proteins belong to a larger group of proteins that are involved in DNA-binding known as sperm nuclear basic proteins (SNBPs). The SNBPs are involved in spermiogenesis and nuclear transformation. Spermiogenesis is the process where round spermatids develop into mature spermatozoa. During spermiogenesis, nuclear transformation occurs where histones are exchanged for protamines, the chromatin condenses, and nuclear shape becomes elongated like a needle in Drosophila. In the current work, we were interested in the role that sperm nuclear basic proteins (SNBPs) play in chromatin condensation and nuclear transformation, and in sperm nuclear shaping during spermatogenesis in Drosophila. Our goal was to search the 12 sequenced Drosophila genomes for SNBPs based on the known SNBP sequences for *D. melanogaster*.

The analysis was initially conducted using the basic local alignment search tool (BLAST) which utilizes a conservative algorithm to compare primary biological sequence information. Searches were performed on genomic DNA, RNA transcripts and amino acid sequences from 12 species of Drosophila flies whose genomes have been sequenced. The best matches from each of the 12 Drosophila species were aligned using CLUSTALW. Sequence alignments and analysis of amino acid content indicate that homologues to Mst35Ba and Mst35Bb are present in all 12 species of flies analyzed in this study. Additionally, a T-Coffee analysis found a conserved region among the isolated sequences that appears to contain a high mobility group DNA binding box. The protein functional domains were found through Domain Annotation –InterPro Scan on

Swiss-MODEL Workspace and Imperial College London Phyre 2. Lastly, Imperial College London Phyre2 tool was used to predict secondary structures.

Preliminary molecular and ultrastructural results were also generated. Genomic DNA from *D. pseudoobscura* was extracted and PCR products were generated based on putative sequence for *D. pseudoobscura* GA18970. Finally, transmission electron microscopy was performed on sperm from *D. pseudoobscura* testes and seminal vesicles, and initial analysis of chromatin condensation patterns was performed.

Introduction

I. Overview

Drosophila melanogaster has been used as a model organism for studies of genetics, evolution, development and cellular biology for the last century (Gilbert et al. 2008). There are currently 12 of the approximately 2800 fly species in the family of *Drosophilidae* that have been sequenced (Markow and O'Grady, 2007; Figs. 1, 2A, and 2B). Following the sequencing of *D. melanogaster*, the second Drosophila fly to be sequenced was *D. pseudoobscura* because of its evolutionary relationship to *D. melanogaster* (Richards et al. 2005; Markow and O'Grady, 2007). This relative of *D. melanogaster* has recently been used for *in vitro* spermatogenesis studies by our lab (Njogu et al. 2010; and Ricketts et al. 2011).

II. Spermatogenesis and Spermiogenesis in Drosophila

The process of mature sperm formation in adult male Drosophila is similar to mammalian spermatogenesis. In flies, spermatogenesis proceeds within blind-ended tubular or ellipsoid testes. It begins in the apex of the testes in a region known as the stem cell niche (White-Cooper et al. 2009; Ricketts et al. 2011; Fig. 3).

During spermatogenesis, the spermatogenic stem cells in the stem cell niche divide to produce another stem cell and a gonialblast cell (Fig. 3). The gonialblast will enter into spermatogenesis while the stem cell will remain in the niche in an undifferentiated state. After several mitotic divisions (five in *D. pseudoobscura*), the cells undergo two meiotic divisions to produce haploid round spermatids. The post-meiotic stage of spermatogenesis that follows is called spermiogenesis. During spermiogenesis, the round spermatids become elongated spermatids due to the growth of the tail. The different stages of spermatogenesis have been characterized in in vitro cell cultures of *D. melanogaster* and *D. pseudoobscura* (Njogu et al.

2010; Noguchi and Miller 2003; Raja et al. 2005; Ricketts et al. 2011). Nuclear transformation is a process that involves histones being exchanged for protamines, chromatin condensation and the transformation of sperm nuclear shape from spherical to an elongated needle-like shape in Drosophila. During this transformation, the chromatin loses its nucleosome organization as somatic histones are exchanged for sperm-specific nuclear basic proteins (SNBPs) (Eirin-Lopez et al. 2006). SNBPs are categorized into three types: protamines (P type), protamine-like proteins (PL type), and histone H1 linker-like proteins (H type) (Eirin-Lopez et al. 2006). PL type and H type proteins have been found in *D. melanogaster* sperm nuclei and are designated Mst35Ba (PL type; ProtA), Mst35Bb (PL type; ProtB), and Mst35f (H type) (Raja et al. 2005).

III. Sperm Nuclear Basic Proteins

As noted above, SNBPs can be divided into three groups: histone group (histone H1 linker-like proteins*)*; protamine-like proteins; and true protamines. The presence of protamine-like proteins and histone H1 linker-like proteins has been well documented in several invertebrate animals such as *Spisula solidissima, Octopus vulgaris* and *Eledone cirrhosa* (Eirin-Lopez et al. 2006; Ausio 1999), as well as vertebrates such as *Dicentrarchus labrax, Mus musculus, Homo sapiens*, and *Rattus norvgicus* (Saperas et al. 1993; Hammoud et al. 2009).

Detailed analysis of SNBPs has shown that true protamines evolved from protamine-like proteins and protamine-like proteins evolved from histones (Balhorn et al. 2007; Eirin-Lopez et al. 2009). The protamine-like proteins generally have high concentrations of basic amino acids such as arginine and lysine with varying degrees of concentration of the other amino acids (Birkhead et al. 2009; Eirin-Lopez 2006). The importance of arginine appears to be that it has more affinity for binding DNA as compared to lysine (Eirin-Lopez et al. 2006b; Kasinsky et al. 2011). These other amino acids in protamine-like proteins include serine and alanine. Likewise, histone H1 linker proteins have similar amino acid content as compared to protamine-like proteins with lysine usually making up the 44% and approximately 8% being arginine (Birkhead et al. 2009; Kasinsky et al. 2011). Similarly, the protamine-like proteins have an approximate concentration of 35 to 50% of lysine and arginine amino acids combined (Eirin-Lopez et al. 2006). In contrast, true protamines are very rich in arginine (Balhorn, et al. 2007). Interestingly, both protamines and protamine-like proteins are fast evolving and highly variable among species (Eirin-Lopez et al. 2011) including those in the same genus (Rooney et al. 2000).

In *D. melanogaster*, the SNBPs are called male specific transcripts (Mst) (Eirin-Lopez et al. 2006b; Tweedie et al. 2009). There are three known male specific transcripts found in the sperm nucleus in *D. melanogaster*: Mst35Ba, Mst35Bb, and Mst77F. Mst35Ba and Mst35Bb have been well documented and characterized as DNA binding proteins (Raja et al. 2005; Dorus et al. 2008). The difference between Mst35Ba and Mst35Bb is only two amino acids with Mst35Ba being 146 amino acids and Mst35Bb being only 144 amino acids. This similarity is due to the duplication event of Mst35Ba to Mst35Bb (Dorus et al. 2008; Raja et al. 2005). The last male specific transcript found in the sperm nucleus in *D. melanogaster* is known as Mst77F, which has been shown to be involved in chromatin condensation and nuclear shaping (Raja et al. 2005). The interaction of these SNBPs in *D. melanogaster* give rise to the chromatin condensation patterns that is likely unique to the *D. melanogaster* sperm nucleus (Birkhead et al 2009; Raja et al. 2005; Rathke et al. 2007).

IV. Current Approach

In the current work, we have used the published sequences for the *D. melanogaster* protamine-like proteins Mst35Ba and Mst35Bb to search the genomes of the 12 sequenced Drosophila species for similar SNBPs. Several bioinformatics tools have been used to find putative DNA and protein sequences among the 12 sequenced Drosophila flies. A BLAST search was conducted to find similar DNA and amino acid sequences in the twelve Drosophila flies. T-Coffee, a local sequence alignment tool, was used find a consensus region within the matches. These matches were then used to generate phylogenetic trees using ClustalW2. Three different DNA binding predicting tools were used on the whole matched protein and the conserved regions. This was followed by a search for functional domains for each of the conserved region and the whole proteins. Lastly, a detailed analysis was conducted on the amino acid content of all the matched proteins and their respective conserved regions. Our results indicate that homologues for Mst35Ba and Mst35Bb are present in all 12 sequences Drosophila species. Additionally, the conserved amino acid sequences corresponded to a known DNA-binding high mobility group (HMG) box. We hypothesize that the rapidly evolving and highly variable protamine-like proteins will give rise to variable chromatin condensation patterns, which in turn will give rise to internal nuclei forces that help generate the species-specific shape of the sperm nucleus. The 12 sequenced genomes in the genus Drosophila present a unique opportunity for a large-scale, fine-grained analysis of the SNBPs and their relationship to chromatin patterning, as well as providing a means to closely analyze the evolution of these proteins.

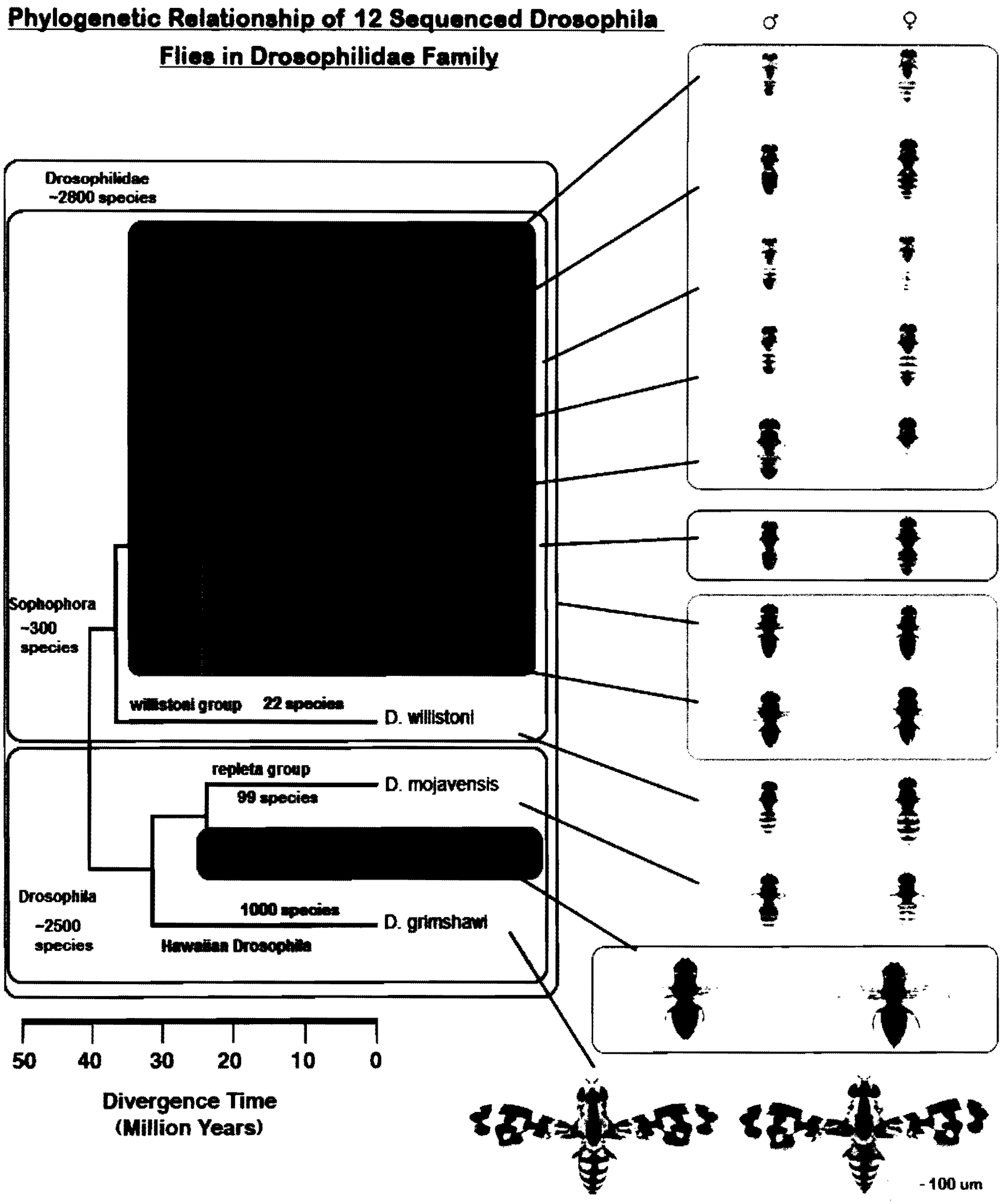


Figure. 1. The phylogenetic relationship among the 12 sequenced Drosophila species in the Drosophilidae family. The time scale illustrates the evolutionary distance in terms of millions of years (modified from Tweedie et al. 2009 and Markow et al. 2002).

Relationship of 12 Sequenced Drosophila Flies within the Drosophilidae Family (Sophophora Subgenus)

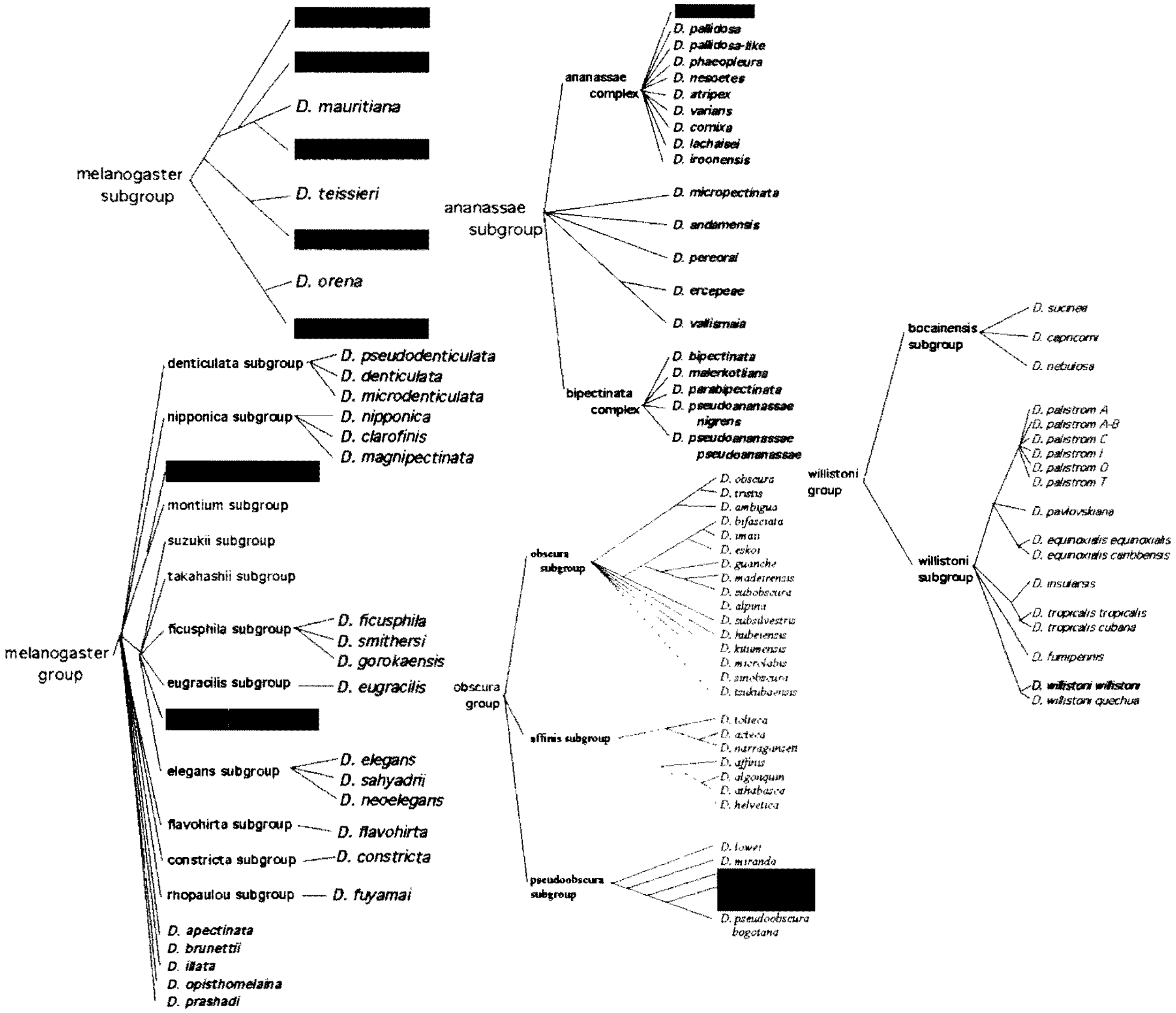


Figure. 2A. Expanded phylogenetic trees of all Drosophila species groups. This figure illustrates the enormous number Drosophila species that are available to be studied within the Drosophilidae Family. Subgenus Sophophora. The highlighted species are the 12 Drosophila species that have been sequenced as shown in Figure 1 (modified from Tweedie et al. 2009).

Relationship of 12 Sequenced Drosophila Flies within the Drosophilidae Family (Drosophila Subgenus)

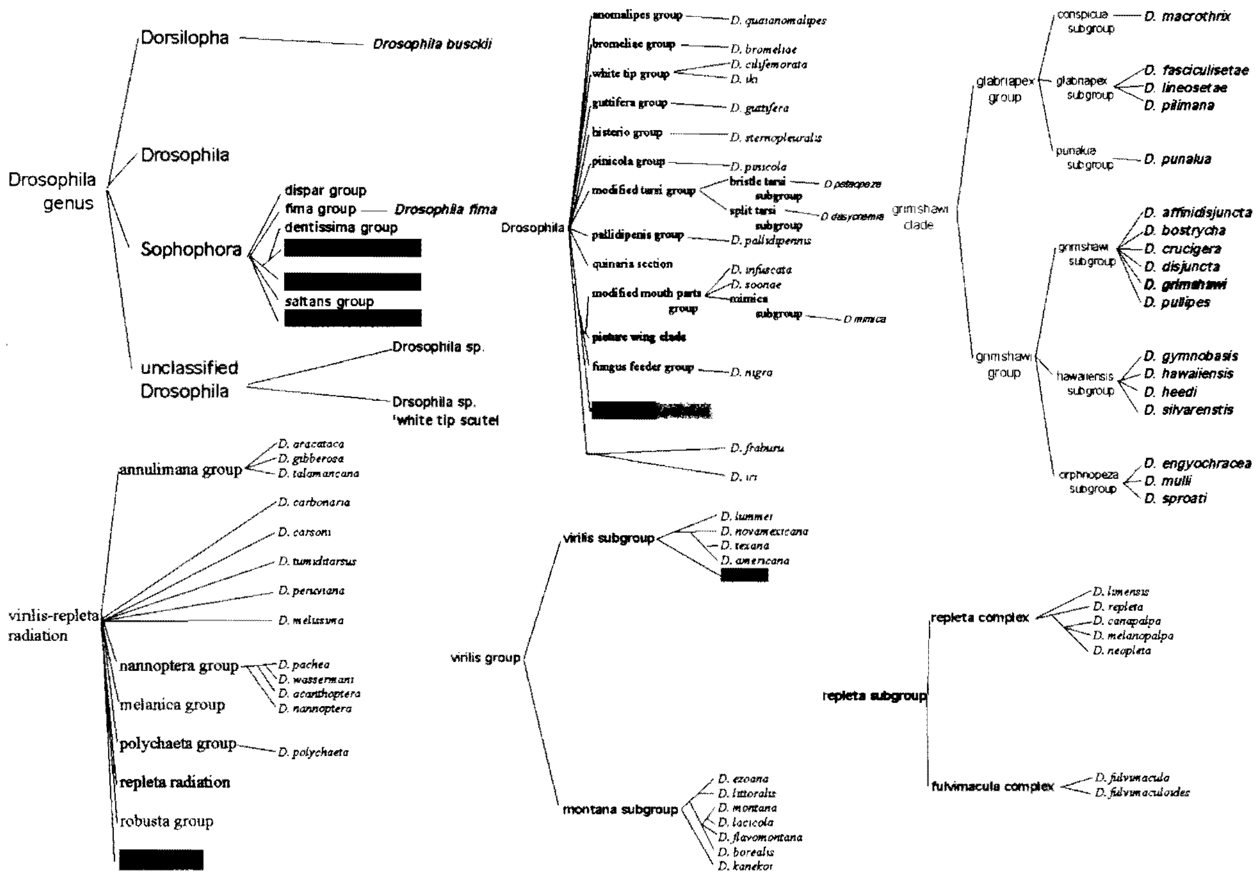


Figure. 2B. Expanded phylogenetic trees of all Drosophila species groups. This figure illustrates the enormous number Drosophila species that are available to be studied within the Drosophilidae Family. Subgenus Drosophila. The highlighted species are the 12 Drosophila species that have been sequenced as shown in Figure 1 (modified from Tweedie et al. 2009).

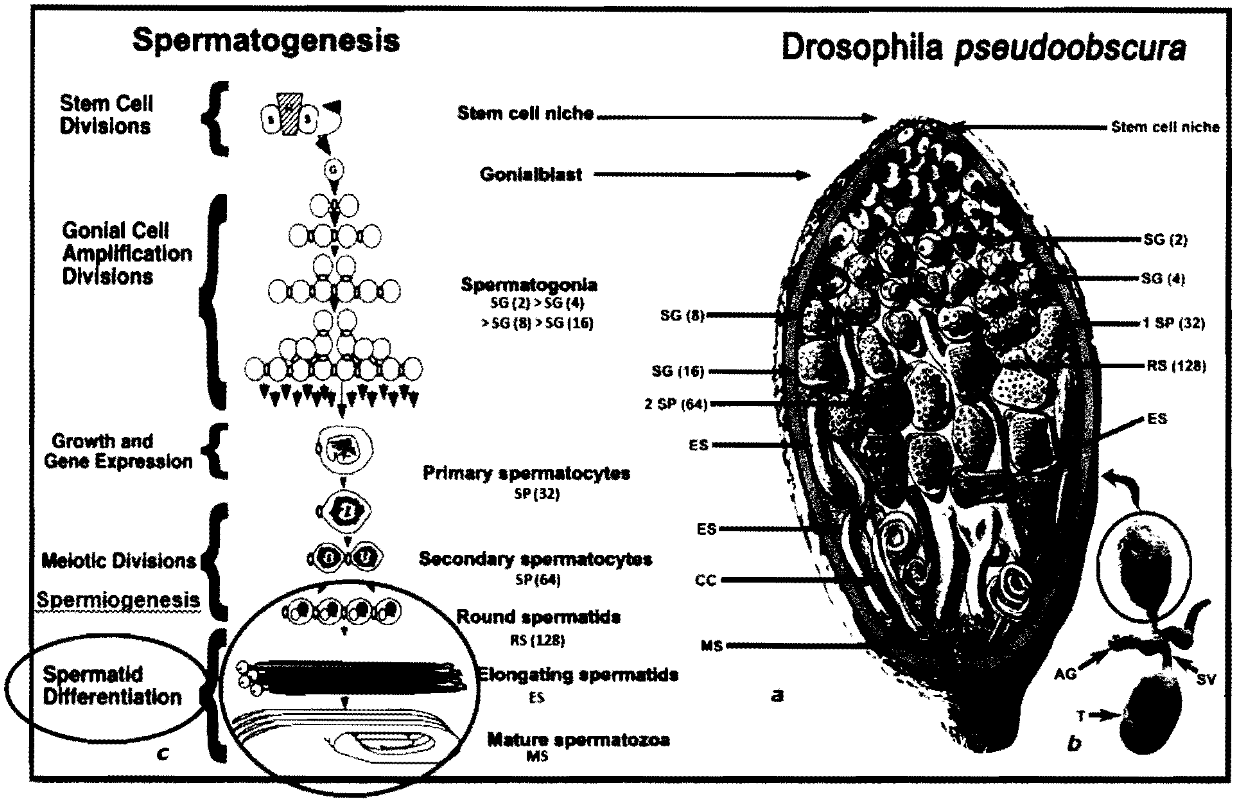


Figure. 3. Illustration of all the stages of spermatogenesis in *Drosophila pseudoobscura.* (a) Illustration of spermatogenic cysts within the testis. H = hub cells, S = somatic stem cells (cyst progenitor cells), G = Glonialblast, SG = spermatogonia, 1° SP = Primary Spermatocyte, 2° SP = Secondary Spermatocyte, SP = Primary spermatocytes, RS = Round spermatids, ES = Elongated Spermatids, MS = Mature spermatozoa, CC = Coiling Cyst. (b) Paired testes, seminal vesicles and accessory glands. AG = Accessory Gland, SV = Seminal Vesicle, T = Testis. (c) Illustration showing cellular changes that occur during spermatogenesis and spermiogenesis. During spermiogenesis, there is growth of the tail, relocation of mitochondria and nuclear transformation of the sperm head. Nuclear transformation entails the histones being exchanged for protamines, chromatin condensation, and the nuclear shape changing from a spherical to an elongated needle like shape in Drosophila (Figure modified from Fuller et al. 1998; Njogu et al. 2010; Zhou et al. 2009; Ricketts et al. 2011).

Methods and Materials

I. Nucleotide BLAST and protein BLAST of ProtA, ProtB in 12 sequenced Drosophila species

The nucleotide reference sequences of the male specific transcripts for *Drosophila melanogaster* protamine-like proteins Mst35Ba (GI: 45549065) and Mst35Bb (GI: 24584359) were obtained through the NCBI nucleotide database. Likewise, the protein reference sequences of the male specific transcripts for *D. melanogaster* protamine-like proteins Mst35Ba (GI: 17137016) and Mst35Bb (GI: 17137018) were obtained through NCBI protein database (NCBI and Clark et al. 2007). Nucleotide BLAST, protein BLAST, and PSI BLAST searches were conducted on the 12 sequenced Drosophila genomes with the respective male specific transcripts from *D. melanogaster* as the controls. The matches were verified and refined through BLASTx and NCBI open reading frame finder (ORF Finder). Subsequently, these matches were aligned with their respective male specific transcript control sequences. The matched nucleotide sequences for the 12 sequenced Drosophila genomes have been listed in Figures 4A, 4B, 5A and 5B. Similarly, the matched protein sequences for the 12 Drosophila genomes have been listed in Figures 6A and 6B. The whole gene regions that correlated to the male specific transcripts for *D. melanogaster* protamine-like proteins Mst35Ba (Flybase ID: FBgn0013300) and Mst35Bb (Flybase ID: FBgn0013301) were obtained through Flybase.org. Then nucleotide BLAST was conducted on 12 sequenced Drosophila flies with respective whole genome matches for *Drosophila melanogaster* male specific transcripts used as controls. The matches were checked with BLASTx.

II. Phylogeny Generation

ClustalW2, a global alignment bio-informatics tool, was used to create phylogenies based on the best matches for each respective nucleotide and protein sequences: (NCBI nucleotide transcript matches, Flybase nucleotide matches, and protein matches).

III. Conserved Regions

T-Coffee, a local alignment bio-informatics tool, was used to find the conserved regions among the best protein matches for their respective male specific transcripts for the 12 sequenced species (http://tcf_dev.vital-it.ch/apps/tcoffee/index.html; Di Tommaso et al. 2011). Partial Order Alignment Visualization (POAVIZ) was used to demonstrate the overall conservation of the transcript mRNA and protein best matches for their respective male specific transcript (Lee et al. 2002; Grasso et al. 2003).

IV. Amino Acid Content Analysis

Sequence Manipulation Suite Protein Statistics and Graphpad Prism 5.0 were used to generate bar graphs and statistically analyze each of the SBNP protein BLAST matches and conserved sequences (http://www.bioinformatics.org/sms2/protein_stats.html). Additional sequences for histone H1 linker like proteins, protamine-like proteins, and true protamines were added to the analysis to illustrate evolutionary relationship of the protein BLAST results. The following histone H1 linker proteins were added: *Mus musculus* spermatid-specific linker histone H1-like protein (GI: 9055232) and *Rattus norvegicus* histone linker H1 domain, spermatid-specific 1 (GI: 157818369). The following protamine-like proteins were added: *Mullus surmuletus* protamine-like protein (GI: 115565002), *Spisula solidissima* sperm nuclear basic protein PL-I isoform PLIa (GI: 48526358), and *Spisula solidissima* sperm nuclear basic protein PL-I isoform PLIb (GI: 48526360). The following true protamines were added: *Homo sapiens* sperm protamine P1 (GI: 4506109), *Homo sapiens* protamine-2 (GI: 68989267), *Mus musculus* sperm protamine P1 (GI: 7305409), *Mus musculus* protamine-2 (GI: 6679475), and *Dicentrarchus labrax* sperm protamine (GI: 263998).

V. Putative DNA Binding Domains

DNA-Binder, BindN+ and BindN-RF were used to generate statistical graphical data to further analyze each SBNP protein BLAST matches and conserved domains found through T-Coffee alignment. DNA-Binder verified the sequences 3 algorithms: realistic, main, and alternative (http://www.imtech.res.in/raghava/dnabinder/; Kumar et al. 2007). BindN+ and BindN-RF showed the actual DNA Binding residues on the protein sequences (http://bioinfo.ggc.org/; Wang et al. 2009; Wang et al. 2010).

VI. Putative 2D Secondary Structure and Protein Disorder Prediction

The putative secondary structures and their protein disorder for each SBNP protein BLAST matches and conserved regions (Putative DNA Binding Domain) were predicted using several bio-informatics tools that yielded similar results. These tools included the following: UCL Psi-Pred (http://bioinf.cs.ucl.ac.uk/psipred/); UCL Diso-Pred (http://bioinf.cs.ucl.ac.uk/disopred); Swiss-Model-Workspace Domain Annotation Tool (http://swissmodel.expasy.org/workspace/; Arnold, et al. 2006); and Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/; Kelley et al. 2009).

VII. Functional Groups, 3D Secondary Structures, and Putative Tertiary Models

The functional groups for each respective SBNP protein BLAST match and conserved regions (putative DNA Binding Domains) was found through Swiss-Model-Workspace Domain Annotation Tool (http://swissmodel.expasy.org/workspace/; Arnold et al. 2006); and Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/; Kelley et al. 2009). Both yielded similar results. The overall functional groups were derived using Swiss-Model-Workspace Domain Annotation Tool. Phyre2 was used to generate the putative 3D secondary and tertiary models of each respective SBNP protein BLAST match and conserved regions (Putative DNA Binding Domains). The 3D secondary structure models were further analyzed through Molsoft ICM Browser (http://www.molsoft.com/).

VIII. Primer Design

Primers were designed for *Drosophila pseudoobscura* gene transcript location of GA18970 (GI: 198475489) using NCBI Primer BLAST and IDT PrimerQuest^{SM} respectively. The primer matches were analyzed through IDT OligoAnalyzer (http://www.ncbi.nlm.nih.gov/tools/primer-blast/; http://www.idtdna.com). The primers are listed in Table 5. Primers were synthesized at MWG Operon (http://www.operon.com) and shipped to our lab.

IX. Fly Stocks and Cultures

Living fly stocks were acquired from the San Diego Drosophila Species stock center and maintained in our lab at room temperature (25°C) on Drosophila Jazz Mix medium (Fisher).

X. DNA Extraction, PCR, and Sequencing

The Qiagen Kit (QIAMp) was used in the extraction of DNA from Drosophila flies followed the manufacturer's protocol (Qiagen, Valencia, CA). Two *D. pseudoobscura* flies were placed in each 1.5 mL centrifuge tube. The flies were then cooled by either placing them on ice or in the freezer for approximately 90 seconds. Fifty micro-liter of ATL Buffer was then added into the centrifuge tube. A combination of "homemade" grinders based on 200 µl pipetman tips and specialized centrifuge grinders were used to grind the flies. Another 50 µL of ATL Buffer was then added. Then 10 µL of proteinase K and 100 µL Buffer AL was added to remove proteins. The samples were vortexed for approximately 15 sec. The ground fly parts were incubated at 56°C for 10 minutes on a heat block or a warm water bath on a rocker. Next, 50 µL of 100 % (200 Proof) ethanol was added to the centrifuge tube, vortexed, and incubated at room temperature for 3 minutes. The centrifuge tubes were quickly centrifuged to pull down any liquid from the lid. The lysate was carefully transferred to a Qiagen MinElute Column that contained. A new collection tube with the column was centrifuged at 8000 rpm for 1 minute. Then 500 μL of AW1 was carefully added so that the rim of MinElute Column was not wet. The column was centrifuged at 8000 rpm for 1 minute. Another collection tube was prepared with 500 μL of Buffer AW2 and added to the column. The new collection tube with the column was centrifuged at 8000 rpm for 1 minute. The column was moved to a new collection tube. A dry centrifugation step at full speed (13,200 RPM) for 3 minutes following in order to dry the membrane on Qiagen MinElute Column, and the collection tube discarded. The column was added to 1.5 mL tube and was incubated at RT for 5 minutes after 20 μL of diH_2O was added to membrane of the Qiagen MinElute Column. After the incubation step, the column with 1.5 mL tube was centrifuged at 13,200 rpm for 1 minute. Twenty micro-liter of diH_2O was added to the center of the membrane and incubated at RT for 5 minutes. The 1.5 mL tube with column was centrifuged for 1 minute at 13,200 rpm. This process yielded approximately 40 μL of extracted DNA. Extracted DNA was analyzed by 1% agarose gels.

Polymerase Chain Reaction (PCR) samples were prepared with each PCR tube containing a total volume of 25 μL with following reagents: 1 μL of extracted DNA from *Drosophila pseudoobscura*, 12.5 μL of Hot-Start Taq Mastermix with 1.5 mM of $MgCl_2$ (Denville Scientific), 2.5 μL dimethyl sulfoxide (DMSO), 7 μL of sterile diH_2O, and 1 μL of the respective forward and reverse primer as shown in Table 12. PCR was conducted with the denaturation stage set to 1 cycle of 95°C for 5 minutes. The annealing stage was set to 35 cycles of 95°Cfor 40 seconds, 60°C for 40 seconds and 72°C for 40 seconds. The elongation stage was set to 72°C for 5 minutes. PCR products were then analyzed by 2% agarose gels. The PCR primers were diluted with sterile diH_2O using a 1:200 ratio, for a final concentration of 500 nM. PCR products were sequenced by Genewiz (South Plainfield, NJ). These results were then analyzed with NCBI nucleotide BLAST2.

XI. Transmission Electron Microscopy

Flies were anaesthetized on ice or using CO_2. Testes were dissected in a drop of 1X PBS and transferred to 2% glutaraldehyde in 0.1 M cacodylate buffer in a spot plate well. Testes were fixed for 1-2 hours at room temperature (RT), and then rinsed two times in 0.1 M cacodylate buffer for 30 minutes on a rotating platform at room temperature. The final rinse in 0.1 M cacodylate buffer was done overnight at 4 degrees C. The following day, the testes were postfixed in 1% osmium tetra-oxide in 0.1 M cacodylate for 1 hour at room temperature. The samples were rinsed three times in 0.1 M cacodylate for 15 minutes at room temperature each rinse. The testes were then dehydrated in an ethanol series: 50%, 70%, 95% (2 times) for 10 minutes each. The final dehydration step was in 100% ethanol, two rinses for 20 minutes each at RT. Samples were transferred to 100% acetone and rinse for 15 minutes. The testes were then infiltrated with Embed 182 resin as follows. Samples were transferred from acetone into a 1:1 resin: acetone and put on a rotator for 1 hour. The mixture was removed and replaced with 2:1 resin: acetone and placed on the rotator overnight. The following day, the samples mixed with 100% resin for at least 1 hour, and then transferred to fresh 100% resin for at least 1 additional hour. The testes were then gently removed from the resin with a sharpened wooden applicator and placed at the bottom of a size 00 plastic BEEM capsules (Electron Microscopy Sciences). The capsules were filled with 100% resin and incubated overnight at 60 degrees C. Ultrathin sections (60-80 nanometers) were made on a Leica ultramicrotome using a diamond knife.

Sections were stained with 1% uranyl acetate, rinsed in distilled water, and viewed on an FEI Tecnai G2 Spirit transmission electron microscope.

Results

I. BLAST results for nucleic acid sequences in 12 Drosophila species

Using the published genomic and mRNA nucleotide sequences for Mst35Ba (GI: 45549065) and Mst35Bb (GI: 24584359) we searched the sequenced genomes for the 12 Drosophila sequenced species. Figure 4A shows all of the Drosophila matches for Mst35Ba: *Drosophila simulans* (GI: 195579289), *Drosophila sechelia* (GI: 195338498), *Drosophila yakuba* (GI: 195474092), *Drosophila erecta* (GI: 194857282), *Drosophila anannassae* (GI: 194758514), *Drosophila pseudoobscura* (GI: 198475489), *and Drosophila persmillis* (GI: 195159817), *Drosophila* willistoni (GI: 195437082), *Drosophila mojavensis* (GI: 195115614) and *Drosophila virilis* (GI: 195385648). Although there were six unique matches for *Drosophila* willistoni, only the best match is shown. Figure 4B shows all the mRNA transcript matches for Mst35Bb. The only difference being *Drosophila grimshawi* (GI: 195043630) is the best match for Mst35Bb, while *Drosophila grimshawi* (GI: 195055896) is the best match for Mst35Ba. Both of the *Drosophila grimshawi* matches were attained through the modification of the nucleotide NCBI search parameters of match/mismatch score set to (4,-5) and the maximum target sequence to be displayed set to 100.

In the melanogaster subgroup, *D. simulans and D. sechelia* transcript (mRNA) matches for both Mst35Ba and Mst35Bb had the identical E-value score of 6e-139, maximum identity of 84% and query coverage of the region of 62%. Their genomic DNA matches had identical maximum identity scores of 100%, E-value scores for only Mst35Ba as illustrated in Figure 5A. As indicated in Figure 5B, the maximum identity score stayed the same except it was slightly reduced to 85%, but the E-value score changed between the two closely related species. As for the genomic E-value scores, the E-value for *D. sechelia* was reduced to 5e-164 and 2e-161 for *D. simulans* respectively. This indicated that the matches for *D. simulans* and *D. sechelia* are closer to Mst35Ba in terms of their genomic DNA relationship. The global ClustalW2 alignment (Figs. 7A and 7B) for the transcripts (mRNA) shows the phylogenetic relationship to be identical to the established phylogenetic tree (Fig. 1). This was further confirmed through the ClustalW2 alignment for the genomic DNA, which shows phylogenetic relationship in Figures 8A and 8B.

The other two Drosophila species in the melanogaster subgroup, *D. erecta* and *D. yakuba*, respectively yielded interesting results in term of their transcripts (mRNA) and their genomic DNA sequence relationship. The mRNA transcript match for *Drosophila yakuba* Mst35Ba and Mst35Bb shows the query coverage was identical for Mst35Ba and Mst35Bb. These types of identical matches are likely to occur due to Mst35Bb were formed as result of a duplication event of Mst35Ba (Raja et al. 2005; Birkhead et al. 2009). Therefore, they are very similar to each other. In Figures 4A and 4B, the *D. yakuba* E-value for Mst35Ba was 8e-57 and 5e-59 for Mst35Bb with the maximum identity for Mst35Ba and Mst35Bb being 84% and 85% respectively. The genomic DNA *D. yakuba* matches are shown in Figures 5A and 5B. The matches for Mst35Ba and Mst35Bb E-values are 5e-54 and 4e-59 and with the maximum identity being 76% and 85% respectively.

D. erecta had interesting results as well with their transcript (mRNA) sequence and genomic DNA matches. In Figures 4A and 4B, the E-value for both Mst35Ba and Mst35Bb match was different with 2e-77 and 2e-83 being the respective scores for each. The maximum identity score for *D. erecta* Mst35Ba and Mst35Bb matches were 81% and 64% respectively for each. As shown in Figures 5A and 5B, the genomic DNA matches for *D. erecta* Mst35Ba and Mst35Bb match had same maximum identity score of 81%. The E-value scores were slightly different with 7e-65 and 3e-74 being the respective scores for *D. erecta* Mst35Ba and Mst35Bb matches. This means that the match for *D. erecta* match is slightly closer to the nucleotide sequence of Mst35Bb when compared to Mst35Ba.

Figures 7A and 7B illustrate that *D. yakuba and D. erecta* are closely related to just each other in their phylogenetic relationship by splintering away from the group formed by *D. melanogaster, D. simulans, and D. sechelia.* Therefore, the melanogaster subgroup is not conserved and is not akin to the established phylogenetic tree. However, as shown in Figures 8A and 8B, the genomic DNA sequences for *D. yakuba* and *D. erecta* illustrate the conservation of melanogaster subgroup, which is akin to the established phylogenetic tree (Fig. 1). Overall, the genomic DNA sequences matches for the different Drosophila species among the melanogaster subgroup are analogous to each other in global manner.

D. annanassae yielded very similar results for its transcript (mRNA) matches for Mst35Ba and Mst35Bb. Figures 4A and 4B depicts this relationship to Mst35Ba and Mst35Bb in terms of its transcript region covered, E-value, and the maximum identity. The E-value scores were 5e-24 and 2e-23 for Mst35Ba and Mst35Bb match for *D. ananassae*. Similarly, the maximum identity was 70% and 71% respectively for Mst35Ba and Mst35Bb with identical query coverage of 28% for these transcript matches.

In Figures 5A and 5B, *D. ananassae* genomic DNA sequence match yielded almost identical results for Mst35Ba and Mst35Bb. The only difference for the genomic DNA sequence was that the Mst35Bb E-value was 8e-16 compared to 7e-13.

Figures 7A and 7B illustrates a generated phylogenetic relationship of the transcript sequences for *D. ananassae.* This generated phylogenetic tree shows that *D. ananassae* does not branch from the melanogaster group. The branching path for *D. ananassae* stems from the initial branched group of *D. erecta and D. yakuba,* which is then followed by the willistoni, repleta, and virilis groups. Figures 8A and 8B, illustrate the genomic DNA phylogenetic relationship and indicates that *D. ananassae* with the obscura group are a third splinter group in the *Drosophilidae* Family. This splinter group is not present in the established phylogenetic tree (Fig. 1) nor are these two flies considered sister species as they do not branch from same branch point. Overall, both of these phylogenetic relationships are drastically different than what is shown in the established phylogenetic tree (Fig. 1). Therefore, there is no Sophophora group present in terms of their global phylogenetic relationship with any of the nucleotide relationships.

D. pseudoobscura and *D. persmillis* (obscura group) matches produced some noteworthy results. Figures 4A and 4B, the transcript (mRNA) matches for *D. pseudoobscura* had similar E-value of 7e-06 and 1e-08 for Mst35Ba and Mst35Bb with the same maximum identity score of 73%. In the case of the *D. persmillis* matches, the results similar with the E-value for Mst35Bb being slightly closer to zero with 2e-09 when compared to Mst35Ba match of 9e-07. The maximum identity score was 73% for both Mst35Ba and Mst35Bb transcript match. Although the query coverage of the transcripts was slightly greater with Mst35Bb 11% compared to Mst35Ba 8% for *D. pseudoobscura* and *D. persmillis*.

Figures 5A and 5B illustrate the obscura group matches for the genomic DNA sequences. The genomic DNA sequences yielded similar results to the transcript sequences with the relationship of the Mst35Bb match being slightly better when compared to Mst35Ba. The E-value for *D. pseudoobscura* for Mst35Bb was 2e-07, which is closer to zero when compared to 1e-04 of Mst35Ba. The closer E-value to zero demonstrates that the match is significant because it is exponentially inversely related to the score of the sequences. Hence, the higher score for the match will yield a lower E-value, which would mean the match is significant. Additionally, the query coverage for *D. pseudoobscuras* Mst35Bb result was 12%, much higher than the 5% coverage for Mst35Ba. The maximum identity for *D. pseudoobscura* and *D. persmillis* was respectively 100% and 73% each for both Mst35Ba and Mst35Bb matches. The E-value for *D. persmillis* was 5e-09 for the Mst35Bb match and 1e-04 for the Mst35Ba match. Also *D. permsillis* query coverage decreased from 5% with Mst35Bb to 2% with Mst35Ba.

Figures 7A and 7B show that the transcript matches for the obscura group branch off from *D. ananassae*, which is similar to the established phylogenetic tree (Fig. 1). Although the branching event in the phylogeny tree to reach the obscura group are different than the established phylogenetic tree (Fig. 1) These differences are expanded upon when examining the phylogenetic relationship of the genomic DNA sequences of *D. pseudoobcura* and *D. persmillis* shown in Figures 8A and 8B. In this case, *D. pseudoobscura* appears to be a sister species to virlis group and *D. persimillis* is a sister species to the Hawaiian Drosophila group. In addition, *D. pseudoobscura* and *D. persmillis* appear to belong to a third group instead of belonging and being a branched group from to *D. ananassae* and within the Sophophora group. These phylogenetic relationship indicate that how protamine-like proteins are very diverse (Fig. 1).

The *D. willistoni* transcript (mRNA) match yielded the same best match for Mst35Ba and mst35bb (Figs. 4A and 4B). The E-value for *D. willistoni* was 1e-22 and a maximum identity score of 100% with query coverage of 25%. The *D. willistoni* genomic DNA sequence matches still had a maximum identity score of 100%, but the E-value score and the query coverage slightly varied (Figs. 5A and 5B). The *D. willistoni* genomic DNA sequence match for Mst35Ba E-value score was 4e-18, which is very close to mst35bb match of 2e-18. The mst35bb match had 10% query coverage and the coverage for Mst35Ba was only 6%.

The overall phylogenetic relationship for the mRNA transcripts is illustrated in Figures 7A and 7B with *D. willistoni* branching further in the beginning from the obscura group and *D. ananassae*. The genomic DNA sequence relationship for *D. willistoni* depicted in Figures 8A and 8B creates a sister specie relationship with *D. ananassae*. In addition, this group is separate from rest of the species in Drosophilidae family.

The repleta and virilis group best matches for *D. mojavensis* and *D. virilis* indicate similar results between their respective transcript (mRNA) matches for Mst35Ba and Mst35Bb as shown in Figures 4A and 4B. The query coverage for *D. mojavensis* was 29%, which is identical for Mst35Ba and mst35bb. The maximum identity score of 67% and the E-value of 9e-13 are slightly better for *D. mojavensis* mst35bb match as compared to the Mst35Ba score of maximum identity of 65% and the E-value of 5e-10. The genomic DNA sequences for Mst35Ba and mst35bb have very similar scores for *D. mojavensis* (Figs 5A and 5B). The E-value for *D. mojavensis* genomic DNA Mst35Ba match is 7e-04 with only 1% query coverage. Whereas the E-value for mst35bb is 4e-05 with a 2% query coverage. In addition, the *D. mojavensis* match for Mst35Ba and mst35bb has 77% maximum identity score.

Comparably, *D. virilis* has similar results with the transcripts (mRNA) for Mst35Ba and Mst35Bb as shown in Figures 4A and 4B. The E-values for *D. virlis* were 2e-07 for Mst35Ba with 10% query coverage and 2e-06 for Mst35Bb with a query coverage of 12% because *D. virilis* is one of the most distantly related specie to *D. melanogaster*. The maximum identity score was 100% for both Mst35Ba and Mst35Bb matches for *D. virilis*, but this time it was for the genomic DNA sequence that matched as shown in Figures 5A and 5B. The E-value scores for the genomic DNA match with *D. virilis* were still very similar with 4e-04 for Mst35Ba and 8e-05 for Mst35Bb. Lastly, the query coverage was smaller as compared to the other sequenced

Drosophila species due to its evolutionary distance.

The results for *D. grimshawi* transcript (mRNA) matches were the same for both Mst35Ba and mst35bb. *D. grimshawi* E-value was 6e-05 and 3% query coverage with 95% maximum identity score was obtained. Figures 5A and 5B show the genomic DNA sequence generated the same maximum identity score of 100% for *D. grimshawi. D. grimshawi* E-value score for Mst35Ba was 7e-04 with a query coverage of 5%. The coverage was increased to 5% for the same gene region, but E-value was reduced to 4e-04 for *D. grimshawi* Mst35Bb match. Whereas E-value for the best match for *D. grimshawi* Mst35bb was 0.001 with only 2% query coverage. The low query coverage is attributed to *D. grimshawi* due to it being the being the evolutionary furthest sequenced fly to *D. melanogaster.*

Figures 8A and 8B shows the transcript (mRNA) phylogenetic relationship for *D. grimshawi* transcript (mRNA) is identical to just the genomic DNA sequences for *D. grimshawi* Mst35Ba match. Whereas the *D. grimshawi* Mst35Bb match, seems to be a sister species to *D. persimilis.* These two species are not sister species in the phylogenetic tree. In the constructed phylogenetic tree *D. grimshawi* branches from the Drosophila subgenus. This type of branching is not present in Figures 8A and 8B, which show *D. grimshawi* branching from the Drosophilidae Family with no relation to the Drosophilia subgenus.

The data shown in Figures 5A and 5B illustrate that majority of the conserved matches occur at the C terminus (3') end of the genomic DNA sequences. Additionally, the phylogenetic relationship of genomic DNA sequences aligns better as compared to the established phylogeny as compared to the transcript (mRNA) phylogenetic alignment (Figs.7A, 7B, 8A and 8B).

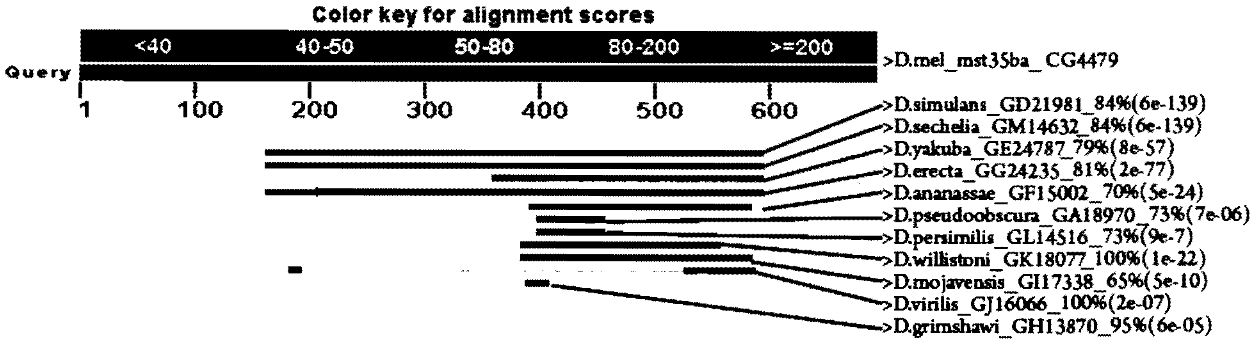


Figure 4A. Nucleotide BLAST of best nucleotide matches among 12 sequenced Drosophila flies for Mst35Ba (ProtA) transcripts (mRNA). Every match was above the threshold and was verified with the NCBI ORF Finder. The exception was the *D. grimshawi* match. The match for *D. grimshawi* was archived through the NCBI nucleotide scoring parameter of match/mismatch scores set to (4,-5) and maximum target sequences to be displayed to 100.

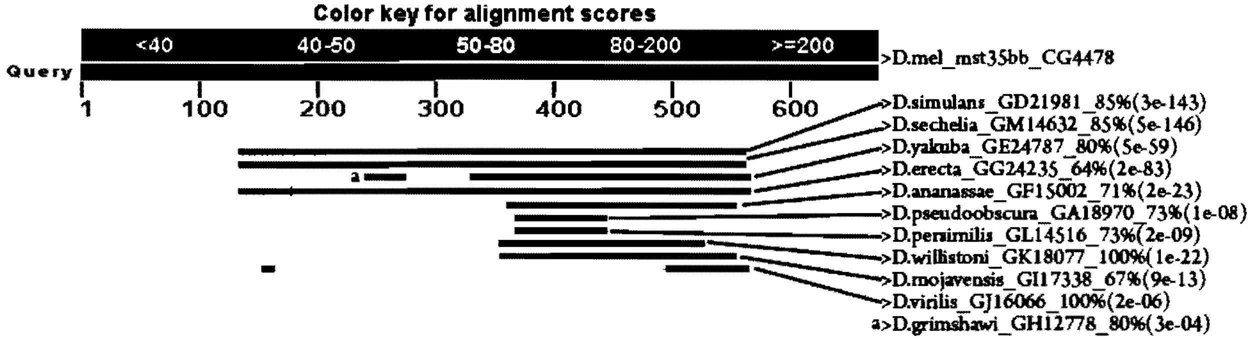


Figure 4B. Nucleotide BLAST of best nucleotide matches among 12 sequenced Drosophila flies for Mst35Bb (ProtB) transcripts (mRNA). Every match was above the threshold and was verified with the NCBI ORF Finder. The exception was the *D. grimshawi* match. The match for *D. grimshawi* was archived through the NCBI nucleotide scoring parameter of match/mismatch scores set to (4,-5) and maximum target sequences to be displayed to 100.

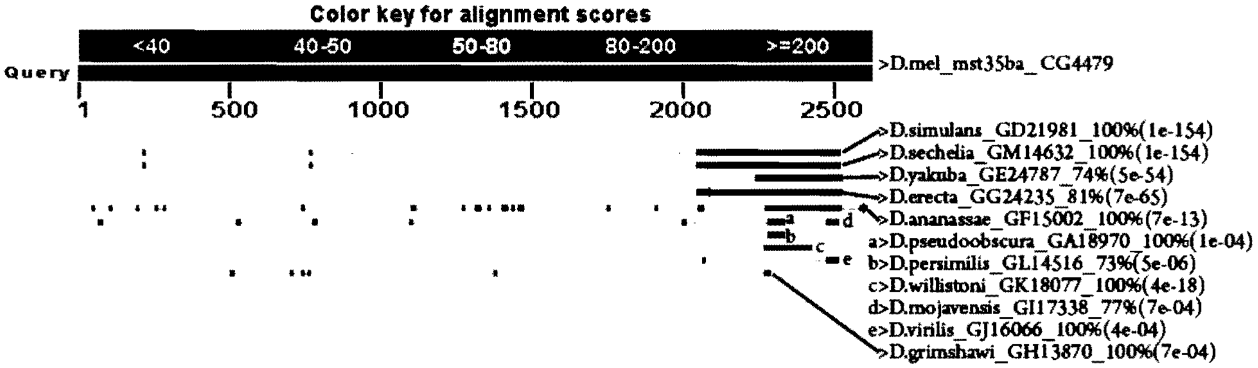


Figure 5A. Nucleotide BLAST of the best Flybase nucleotide (genomic DNA) matches among 12 sequenced Drosophila flies for *D. melanogaster* of Mst35Ba

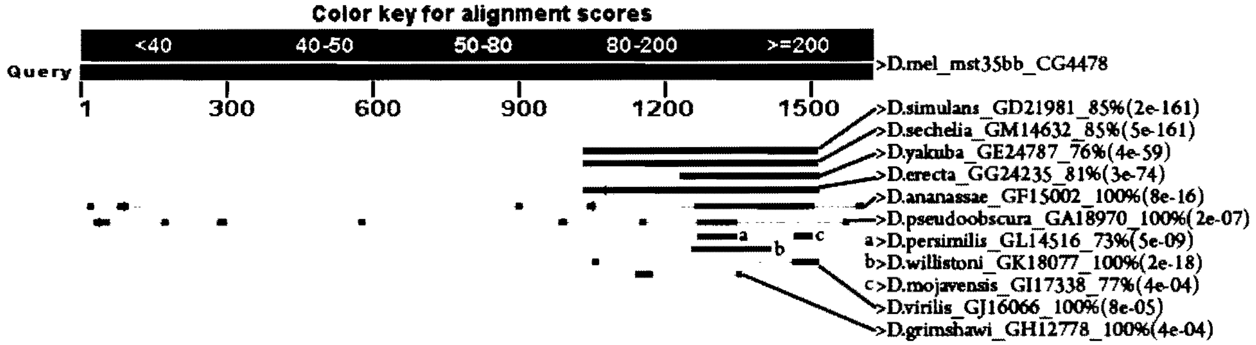


Figure 5B. Nucleotide BLAST of the best Flybase nucleotide (genomic DNA) matches among 12 sequenced Drosophila flies for *D. melanogaster* of Mst35Bb.

Figure 7A. Phylogeny based on NCBI transcript nucleotide (mRNA) best matches of Mst35Ba (Prot A)

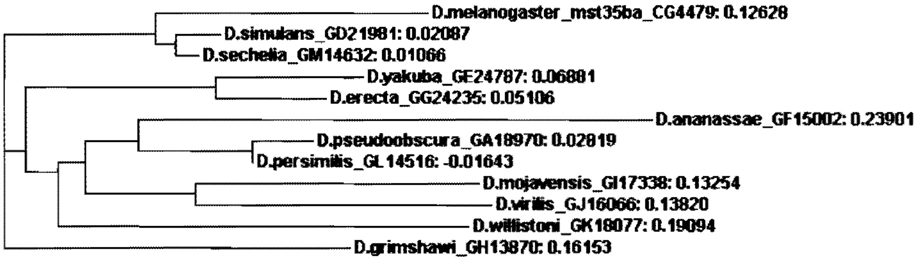
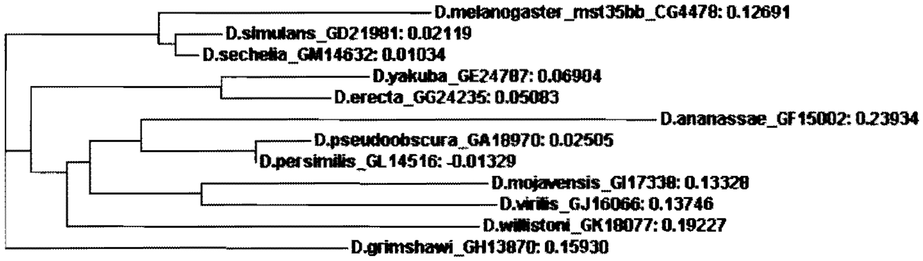

D.melanogaster_mst35ba_CG4479: 0.12628
D.simulans_GD21981: 0.02087
D.sechelia_GM14632: 0.01066
D.yakuba_GE24787: 0.06881
D.erecta_GG24235: 0.05106
D.ananassae_GF15002: 0.23901
D.pseudoobscura_GA18970: 0.02819
D.persimilis_GL14516: -0.01643
D.mojavensis_GI17338: 0.13254
D.virilis_GJ16066: 0.13820
D.willistoni_GK18077: 0.19094
D.grimshawi_GH13870: 0.16153

Figure 7B. Phylogeny based on NCBI transcript nucleotide (mRNA) best matches of Mst35Bb (Prot B)


D.melanogaster_mst35bb_CG4478: 0.12691
D.simulans_GD21981: 0.02119
D.sechelia_GM14632: 0.01034
D.yakuba_GE24787: 0.06904
D.erecta_GG24235: 0.05083
D.ananassae_GF15002: 0.23934
D.pseudoobscura_GA18970: 0.02505
D.persimilis_GL14516: -0.01329
D.mojavensis_GI17338: 0.13328
D.virilis_GJ16066: 0.13746
D.willistoni_GK18077: 0.19227
D.grimshawi_GH13870: 0.15930

The matches for the protamine-like proteins (Mst35Ba and Mst35Bb) are diverse except for *D. simulans* and *D. sechelia* relationship to *D. melanogaster* in the established phylogenetic tree (Fig. 1).

Figure 8A. Phylogeny based on Flybase genomic DNA best matches of Mst35Ba (Prot A)

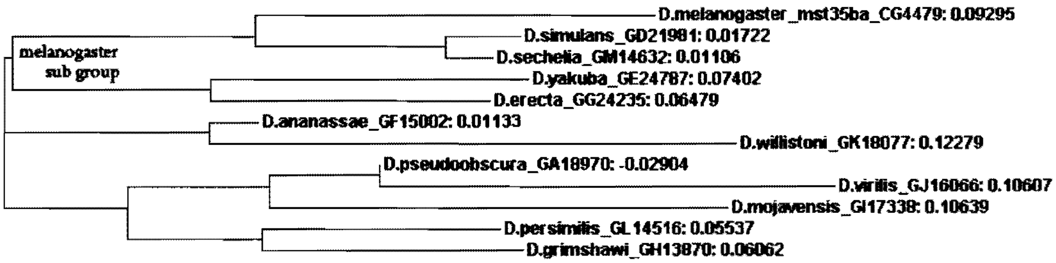
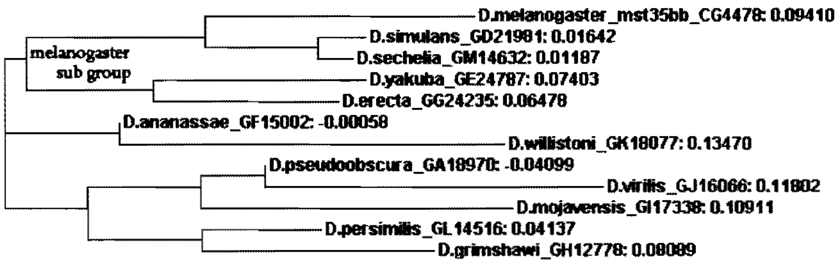

melanogaster sub group
D.melanogaster_mst35ba_CG4479: 0.09295
D.simulans_GD21981: 0.01722
D.sechelia_GM14632: 0.01106
D.yakuba_GE24787: 0.07402
D.erecta_GG24235: 0.06479
D.ananassae_GF15002: 0.01133
D.willistoni_GK18077: 0.12279
D.pseudoobscura_GA18970: -0.02904
D.virilis_GJ16066: 0.10607
D.mojavensis_GI17338: 0.10639
D.persimilis_GL14516: 0.05537
D.grimshawi_GH13870: 0.06062

Figure 8B. Phylogeny based on Flybase genomic DNA best matches of Mst35Bb (Prot B)


melanogaster sub group
D.melanogaster_mst35bb_CG4478: 0.09410
D.simulans_GD21981: 0.01642
D.sechelia_GM14632: 0.01187
D.yakuba_GE24787: 0.07403
D.erecta_GG24235: 0.06478
D.ananassae_GF15002: -0.00058
D.willistoni_GK18077: 0.13470
D.pseudoobscura_GA18970: -0.04099
D.virilis_GJ16066: 0.11802
D.mojavensis_GI17338: 0.10911
D.persimilis_GL14516: 0.04137
D.grimshawi_GH12778: 0.08089

The association of the whole genome nucleotide region in comparison to the transcript region illustrates that the nucleotide matches for protamine-like proteins is akin to the established phylogenetic tree (Fig. 1) for the melanogaster subgroup.

II. Analysis of putative protamine-like proteins in 12 Drosophila species

Using the published protein sequences for Mst35Ba (GI: 17137016) and Mst35Bb (GI: 17137018) we searched the sequenced genomes for the 12 Drosophila species. Figure 6A shows all the Drosophila matches for Mst35Ba: *D. simulans* (GI: 195579290), *D. sechelia* (GI: 195338499), *D. yakuba* (GI: 195474093), *D. erecta* (GI: 194857283), *D. anannassae* (GI: 194758515), *D. pseudoobscura* (GI: 198475490), *and D. persmillis* (GI: 195159818), *Drosophilla willistoni* (GI: 195435143), *D. mojavensis* (GI: 195115615), *D. virilis* (GI: 195385649), and *D. grimshawi* (GI: 195092814). These matches were verified with NCBI ORF Finder, PSI BLAST and protein BLAST. Figure 4B indicates that all the protein matches for Mst35Bb with the only difference being with a different *D. willistoni* (GI: 195437083) match. Although there were six unique matches for *D.* willistoni for Mst35Ba and Mst35Bb, only the best match has been shown. Additionally, all species had a minimum of two matches that were above the threshold except for *D. erecta* that had only match for both Mst35Ba and Mst35Bb.

The best match for each species was re-aligned with its respective control protein sequence for Mst35Ba or Mst35Bb to yield query a coverage percent and an E-value score. All of the species that belonged in the melanogaster subgroup had query coverage of 97%; except for *D. erecta* had query coverage of 88% for the Mst35Ba matches (Figs. 6A and 6B). The query coverage was increased to 99% for all species in the melanogaster subgroup except for *D. erecta* whose query coverage became 89% query for the Mst35Bb match. The query coverage slightly increased for *D. ananassae* between the Mst35Ba coverage of 65% to Mst35Bb coverage of 66% for the same match.

The species that belonged to the obscura group had a greater area of coverage for Mst35Ba as compared to Mst35Bb. As shown in Figures 6A and 6B, the coverage area for *D. pseudoobscura* was 66% for Mst35Ba as compared to 52% for Mst35Bb. Likewise, *D.*

persmillis had coverage of 39% for Mst35Ba, which is 5% larger to the coverage of Mst35Bb (34%). The second best match for *D. pseudoobscura* (GI: 198476418) yielded the same cover area of 43% for Mst35Ba and Mst35Bb. Additionally, the E-value for the second best match for *D. pseudoobscura* Mst35Ba was 7e-14 and 4-e-14 for Mst35Bb. The second best match will be analyzed later with the functional groups of putative conserved regions among the best matches.

D. willistoni generated different best matches for Mst35Ba and Mst35Bb as illustrated in Figures 6A and 6B with the same E-value. The query coverage different with 67% for Mst35Ba match and 75% query coverage for Mst35Bb match.

As shown in Figures 6A and 6B, only *D. mojavensis* for the Drosophila sub genus had the same query coverage of 62% for the Mst35Ba and Mst35Bb match. The query coverage for *D. virilis* for Mst35Bb is larger with 79% than query coverage of 63% match for Mst35Ba. Likewise, the Hawaiian *D. grimshawi* has larger query coverage of 77% for Mst35Bb as compared to Mst35Ba query coverage of 69%.

Figures 6A and 6B illustrate that regardless of the Mst35Ba and Mst35Bb respective best matches; all of the E-values are the same. These E-values range from 2e-28 for *D. pseudoobscura* to 6e-60 for *D. sechelia*. Overall, these E-values are a good indicator that the matches are conserved in terms of their respective control SBNP (Mst35Ba and Mst35Bb).

The global alignment tool ClustalW2–was used to generate phylogenies of the best protein matches. These phylogenies were compared to the generated phylogenetic tree. Figures 9A and 9B show the melanogaster subgroup as being conserved, as with the genomic DNA nucleotide sequences phylogenetic relationship (Figs. 8A and 8B) and the proposed phylogenetic tree (Fig. 1). Another similarity was that *D. ananassae* branched off from the melanogaster sub-group. In the proposed phylogenetic tree, the obscura group branched off *D. ananassae;* the obscura group branched from *D. ananassae* at the same point in the protamine-like protein matches of Mst35Ba shown in Figure 9A. Similarly, the phylogenetic relationship for *D. willistoni* is similar to the known phylogenetic tree (Fig. 1). As shown in Figure 9B, *D. willistoni* branches off from the melanogaster subgroup instead of *D. ananassae* and the obscura group. Also similar to the genomic DNA phylogenetic relationship shown in Figures 8A and 8B, *D. ananassae* and the obscura group are paired together for Mst35Bb matches. The repleta group (*D. mojavensis)* and the virilis group are still paired up together (Figs. 9A and 9B), which is similar to the known phylogenetic tree. The similarity continues with the Hawaiian Drosophila, *D. grimshawi*, branching off from repleta and virilis groups. Whereas in Figure 9A, the *D. grimshawi* branches from the *Drosophilidae* Family instead of branching from the Drosophila sub genus. Overall, the melanogaster subgroup is conserved according to phylogenetic relationship of the putative protamine-like protein matches among the 12 sequenced Drosophila species.

The best protein matches for Mst35Ba and Mst35Bb were statistically analyzed for amino acid percentage the total number of amino acids present in each matched sequence. Additionally the same break down is shown for histone H1 linker-like proteins, protamine-like proteins, and true protamines to illustrate the evolution of histone H1 linker-like proteins to protamine-like proteins and finally to true protamines (Figs. 11A and 11B).

As shown in Figure 11B, the overall number of amino acids for the sister species to *D. melanogaster* Mst35Ba and Mst35Bb are almost identical in the total number of amino acids. The total number of amino acids changes fluctuates between the other matches for Mst35Ba and Mst35Bb, especially for the best match for D. *pseudoobscura* (569 amino acids). A percentage bar graph and table were generated to examine the amino acid content of each protein match

(Fig. 11A and Table 1A). As indicated in previous studies the protamine-like proteins are rich in lysine (K), arginine (R), with a mixture of many other amino acids such as alanine (A), serine (S), and cysteine (C) (N. Saperas; Eirin-Lopez et al. 2006; Birkhead et al. 2009; Kasinsky et al. 2011). Nearly all matched sequences contained substantial amounts of arginine and lysine amino acids (Fig. 11A). Table 1A shows the percentage of all amino acid ratios in matched species. The estimated total percentage of arginine and the lysine amino acids is important for protamine-like proteins. The lowest percentage sum of 13.7% of arginine and lysine were found in the best match for *D. pseudoobscura*. The second best match for *D. pseudoobscura* shows the lysine and arginine combined percentage as 17.4% (Table 1A). This percentage is closer to combined percentage of arginine and lysine for the controls (Mst35Ba and Mst35Bb) with their percentage being 26.7 and 25.7. Also, the melanogaster subgroup arginine and lysine combined ratio is very close to the control sequence with the range being from 25.8 for *D. simulans* to 21.8 for *D. erecta.* (Table 1A) For *D. ananassae*, the second best match had a lysine and arginine combined of 23.2% and the best match percentage was 17.8% (Table 1A). The *D. virilis* combined percentage for arginine and lysine for the best match was 30.9% and 18.5% for the second best match. The Mst35Ba best match for *D. willistoni* had a combined percentage for lysine and arginine of 26.3% and Mst35Bb best match percentage was 31.9%. The *D. mojavensis* combined percentage for arginine and lysine was 16.6%. The *D. virilis* combined percentage for arginine and lysine was 26.8%. Lastly, there was a substantial concentration of serine and various significant concentrations of cysteine present in all matches (Fig. 11A and Table 1A).

The local alignment tool T-Coffee was used to align all the best matches for each SBNP (Mst35Ba and Mst35Bb). A local alignment tool searches for the nucleotide next to it in the relation as compared to a global alignment tool algorithm that searches for the overall consensus.

As shown in Figures 10A and 10B, the overall consensus among each of the best protein matches was very closely aligned. The Mst35Ba consensus was 82% among the best matches and Mst35Bb consensus was 83% among the best matches. In these consensus regions, a conserved region of approximately 56 amino acids for Mst35Ba was found (Fig. 10A). Likewise, a slightly larger conserved region was found in Mst35Bb with its size ranging from 55 to 56 amino acids (Fig. 10B). Then these matches were further analyzed in terms of the amino acid percentage breakdown in Figures 12A and 12B. Also, Tables 1B and 1C illustrated the detailed breakdown of the percentage of amino acids present for each conserved region when compared to the controls and other species.

The average concentration of lysine and arginine for Mst35Ba and Mst35Bb is 26.21%. Whereas as the average concentration of lysine and arginine for the conserved regions for Mst35Ba and Mst35Bb are 25.58%. The average concentration of lysine and arginine in the conserved Mst35Ba and Mst35Bb regions are respectively 25.65% and 25.45%. The best match for *D. pseudoobscura* total percentage of lysine and arginine is 26.79. (Figs. 12A, 12B, Tables 1B and 1C) Whereas the second best match for *D. pseudoobscura* total percentage of lysine and arginine is only 20% with 1.85% of serine amino acids. (not shown) While in the best matches for the rest of conserved regions, once again there appears to be a substantial amount of serine present in the conserved. (Figs. 12A and 12B)

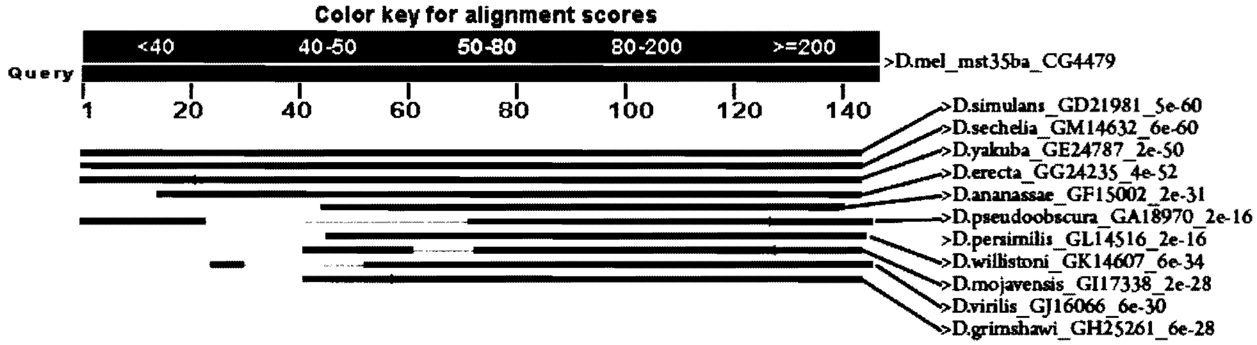


Figure 6A. Protein BLAST of best protein matches among 12 sequenced Drosophila flies for Mst35Ba (ProtA). All of the matched sequences in all twelve sequenced Drosophila flies were based on *D. melanogaster* Mst35Ba. These matches were verified with NCBI ORF Finder, PSI-BLAST, and protein BLAST. The best-matched sequences were aligned through protein BLAST E-values are shown.

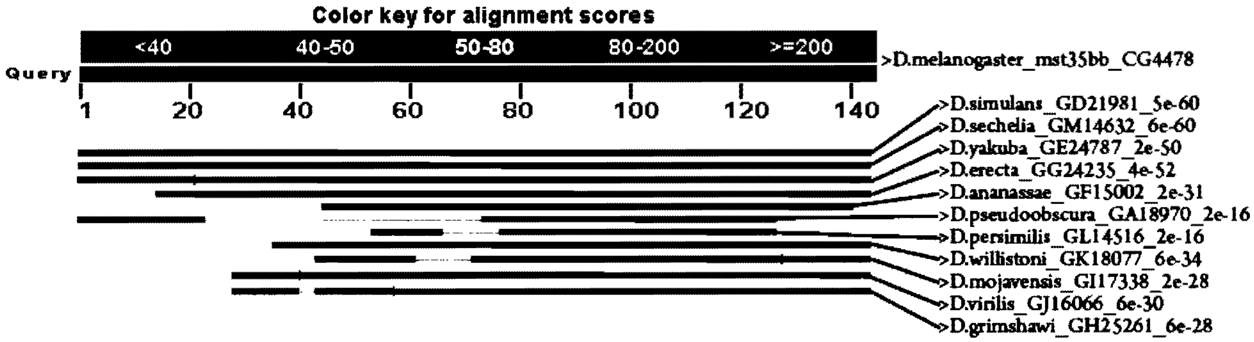


Figure 6B Protein BLAST of best protein matches among 12 sequenced Drosophila flies for Mst35Bb (ProtB). All of the matched sequences in all twelve sequenced Drosophila flies were based on *D. melanogaster* Mst35Bb. These matches were verified with NCBI ORF Finder, PSI-BLAST, and protein BLAST. The best-matched sequences were aligned through protein BLAST E-values are shown.

Figure 9A. Phylogeny based on best protein sequence matches of Mst35Ba (Prot A) with distances

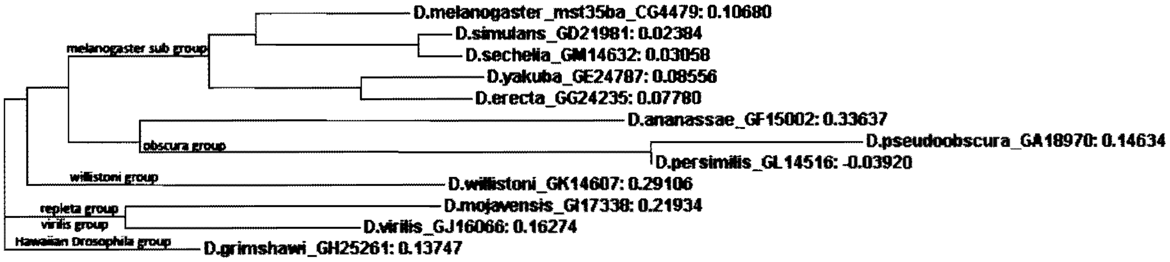
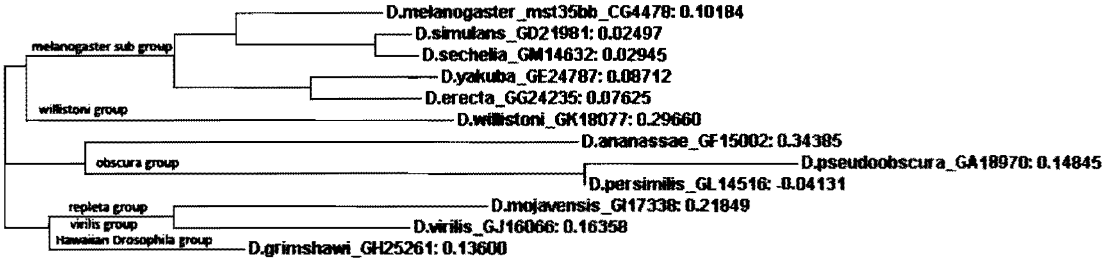
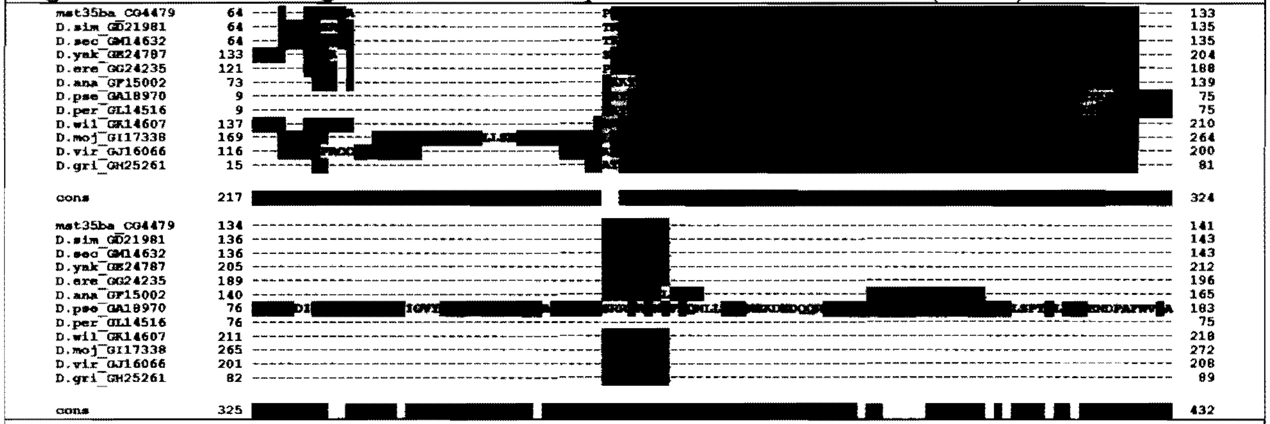

D.melanogaster_mst35ba_CG4479: 0.10680
D.simulans_GD21981: 0.02384
D.sechelia_GM14632: 0.03058
D.yakuba_GE24787: 0.08556
D.erecta_GG24235: 0.07780
D.ananassae_GF15002: 0.33637
D.pseudoobscura_GA18970: 0.14634
D.persimilis_GL14516: -0.03920
D.willistoni_GK14607: 0.29106
D.mojavensis_GI17338: 0.21934
D.virilis_GJ16066: 0.16274
D.grimshawi_GH25261: 0.13747
melanogaster sub group
obscura group
willistoni group
repleta group
virilis group
Hawaiian Drosophila group

Figure 9B. Phylogeny based on best protein sequence matches of Mst35Bb (Prot B) with distances


D.melanogaster_mst35bb_CG4478: 0.10184
D.simulans_GD21981: 0.02497
D.sechelia_GM14632: 0.02945
D.yakuba_GE24787: 0.08712
D.erecta_GG24235: 0.07625
D.willistoni_GK18077: 0.29660
D.ananassae_GF15002: 0.34385
D.pseudoobscura_GA18970: 0.14845
D.persimilis_GL14516: -0.04131
D.mojavensis_GI17338: 0.21849
D.virilis_GJ16066: 0.16358
D.grimshawi_GH25261: 0.13600
melanogaster sub group
willistoni group
obscura group
repleta group
virilis group
Hawaiian Drosophila group

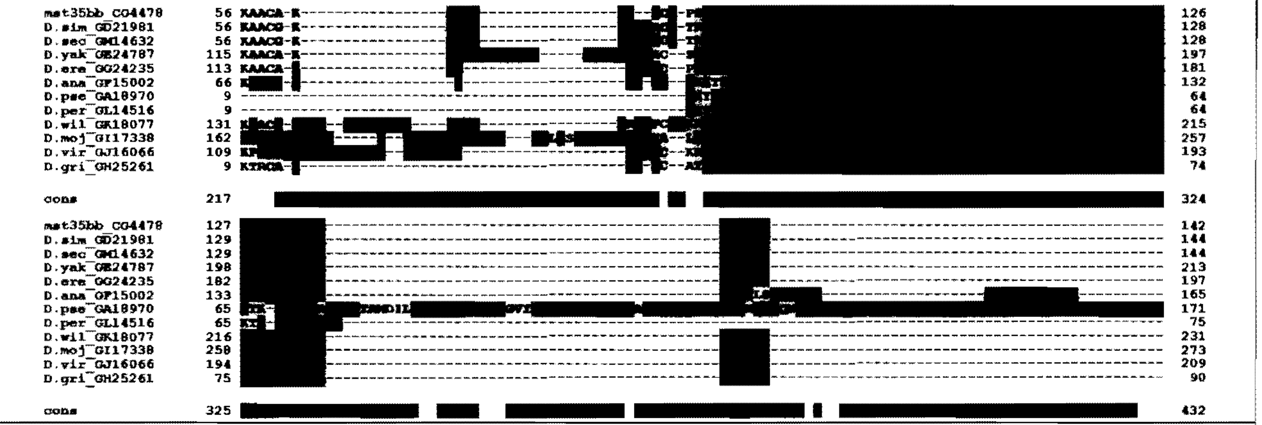
The melanogaster sub-group is conserved among all best protein matches in their respective male specific transcripts. The protein matches for the protamine-like proteins (Mst35Ba and Mst35Bb) indicate that only the melanogaster subgroup is the same as the established phylogenetic tree. (see Fig. 1).

Figure 10A. T-Coffee alignment based on best protein matches of Mst35Ba (Prot A)



The overall consensus for the best matched sequences for Mst35Ba (ProtA) was found to be 82%.

Figure 10B. T-Coffee alignment based on best protein matches of Mst35Bb (ProtB)



The overall consensus for the best matched sequences for Mst35Bb (ProtB) was found to be 83%

Key:

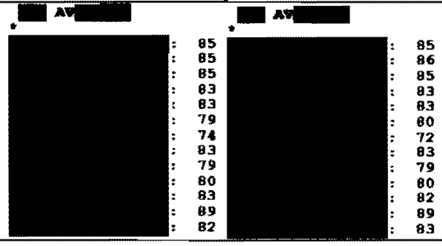


Figure 11A. Amino Acid percentage versus Species (Mst35Ba and Mst35Bb):

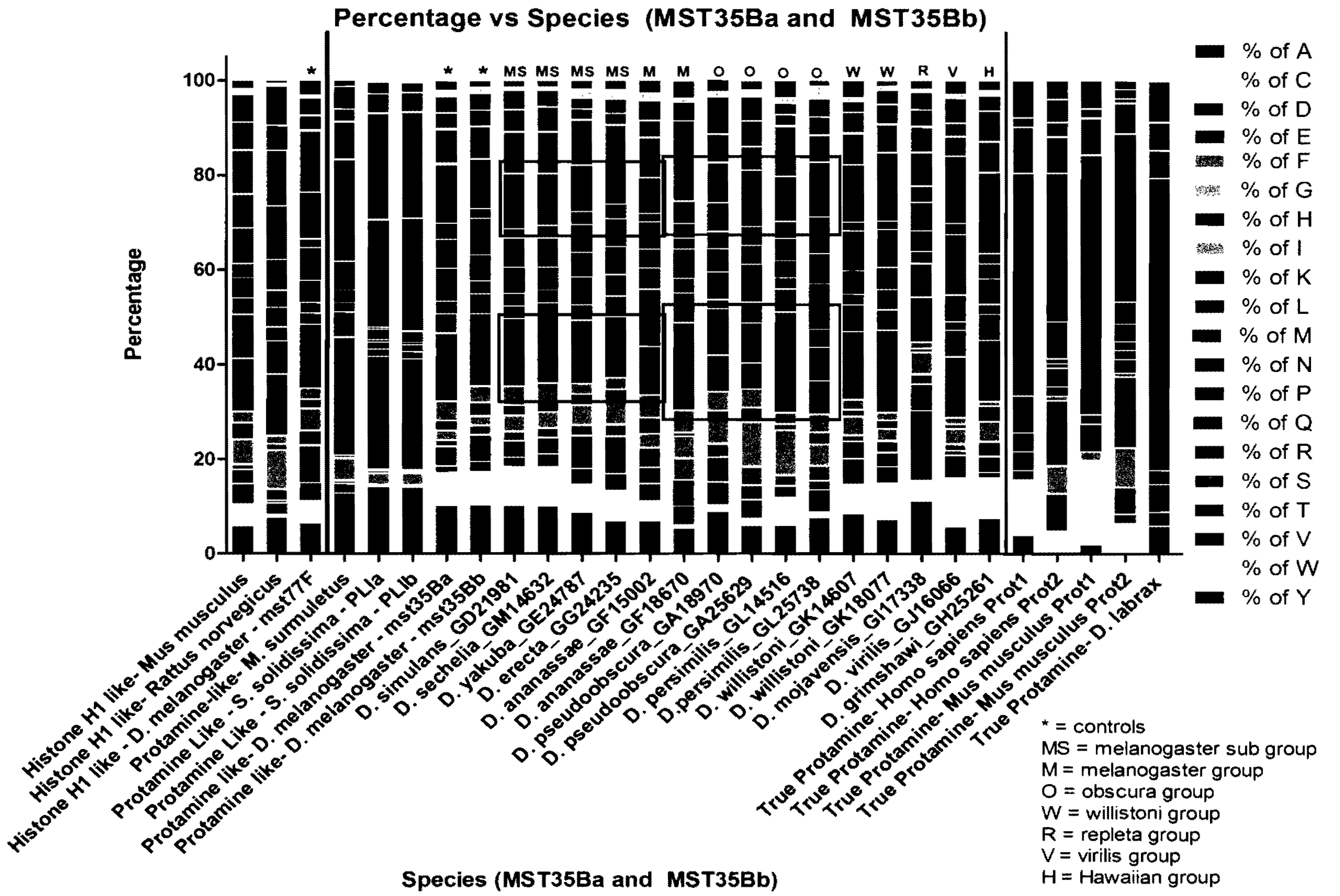


Table 1A. Amino Acid percentage versus Species (Mst35Ba and Mst35Bb)

Species	Percentage of Amino Acids													
	% A	% C	% E	% G	% H	% K	% L	% N	% P	% Q	% R	% S	% T	% V
Histone H1 like- Mus musculus	5.88	4.71	2.94	5.29	3.53	11.18	9.41	4.12	2.94	7.65	7.06	9.41	5.88	5.88
Histone H1 like- Rattus norvegicus	7.69	0.59	0.59	8.28	1.18	13.02	7.1	4.14	3.55	6.51	11.24	11.83	5.33	8.28
Histone H1 like - D. melanogaster - mst77F	6.51	4.65	7.91	4.65	1.86	13.49	2.33	4.65	6.98	1.86	9.77	13.02	3.26	3.72
Protamine-like- M. surmuletus	12.75	0	0	4.7	0	24.83	5.37	2.68	6.04	0	21.48	8.05	2.68	4.7
Protamine-like - S. solidissima - PLla	14.1	0.22	0	2.42	0.44	23.79	1.54	0.44	2.2	0.44	22.69	22.69	3.96	2.42
Protamine-like - S. solidissima - PLlb	14.07	0.22	0	2.42	0.44	23.3	1.54	0.44	2.2	0.22	23.74	22.64	3.96	2.2
Protamine-like- D. melanogaster - Mst35Ba	10.27	6.85	4.11	2.05	2.05	14.38	4.11	6.85	6.16	3.42	12.33	7.53	3.42	3.42
Protamine-like- D. melanogaster - Mst35Bb	10.42	6.94	5.56	2.08	2.78	15.28	4.17	5.56	7.64	2.08	10.42	6.94	3.47	3.47
D. simulans_GD21981	10.2	8.16	3.4	3.4	2.04	14.29	2.72	5.44	6.12	2.04	11.56	8.84	4.76	4.08
D. sechelia_GM14632	10.2	8.16	3.4	3.4	2.04	14.29	2.72	4.76	6.12	2.72	10.88	8.84	5.44	3.4
D. yakuba_GE24787	8.76	5.99	6.91	4.15	2.3	13.36	3.23	5.07	5.99	2.76	9.68	9.68	2.3	2.3
D. erecta_GG24235	6.93	6.44	7.92	4.95	2.48	12.87	4.46	5.94	5.94	2.97	8.91	7.92	2.48	2.97
D. ananassae_GF15002	6.95	4.23	3.32	3.02	3.32	10.27	12.08	3.93	5.74	1.81	7.55	8.46	3.63	4.23
D. ananassae_GF18670	5.43	0.78	5.43	4.65	0.78	18.6	6.2	6.2	2.33	3.1	4.65	9.3	7.75	3.88
D. pseudoobscura_GA18970	8.96	1.41	5.27	5.1	1.93	7.73	9.84	5.8	4.92	3.69	5.98	8.61	5.98	7.91
D. pseudoobscura_GA25629	5.95	1.49	2.99	9.45	2.99	5.47	8.46	7.96	3.98	3.98	11.94	4.98	5.47	4.98
D. persimilis_GL14516	5.95	5.95	2.38	9.52	1.19	21.43	4.76	3.57	4.76	3.57	9.52	10.71	2.38	2.38
D.persimilis_GL25738	7.69	1.28	1.92	4.49	2.56	7.05	7.05	9.62	6.41	7.69	11.54	5.77	3.85	3.85
D. willistoni_GK14607	8.48	6.25	3.57	4.02	1.34	14.29	5.8	4.91	8.48	1.79	12.05	6.7	4.46	2.23
D. willistoni_GK18077	7.32	7.66	2.98	2.55	1.7	17.45	4.26	4.68	8.09	1.7	14.47	5.96	4.26	2.98
D. mojavensis_GI17338	11.19	4.33	5.42	4.69	0.72	9.39	7.22	4.33	6.14	3.25	7.22	5.42	3.61	3.61
D. virilis_GJ16066	5.66	10.38	0.94	3.3	0.94	12.74	5.66	5.66	12.74	2.36	14.15	4.25	2.83	5.19
D. grimshawi_GH25261	7.53	8.6	3.23	4.3	3.23	12.9	5.38	5.38	3.23	2.15	17.2	6.45	6.45	3.23
True Protamine- Homo sapiens Prot1	3.92	11.76	0	0	1.96	0	0	0	3.92	7.84	47.06	9.8	1.96	0
True Protamine- Homo sapiens Prot2	0	4.9	7.84	5.88	13.73	1.96	3.92	0	0.98	7.84	31.37	7.84	2.94	4.9
True Protamine- Mus musculus Prot1	1.96	17.65	0	0	0	5.88	0	0	0	0	54.9	7.84	1.96	0
True Protamine- Mus musculus Prot2	0	6.54	5.61	8.41	14.95	2.8	1.87	0	3.74	4.67	35.51	6.54	0.93	1.87
True Protamine- D. labrax	5.88	0	2.94	0	0	0	0	0	5.88	2.94	61.76	5.88	5.88	8.82

Figure 11B. Number of Amino Acids versus Species (Mst35Ba and Mst35Bb)

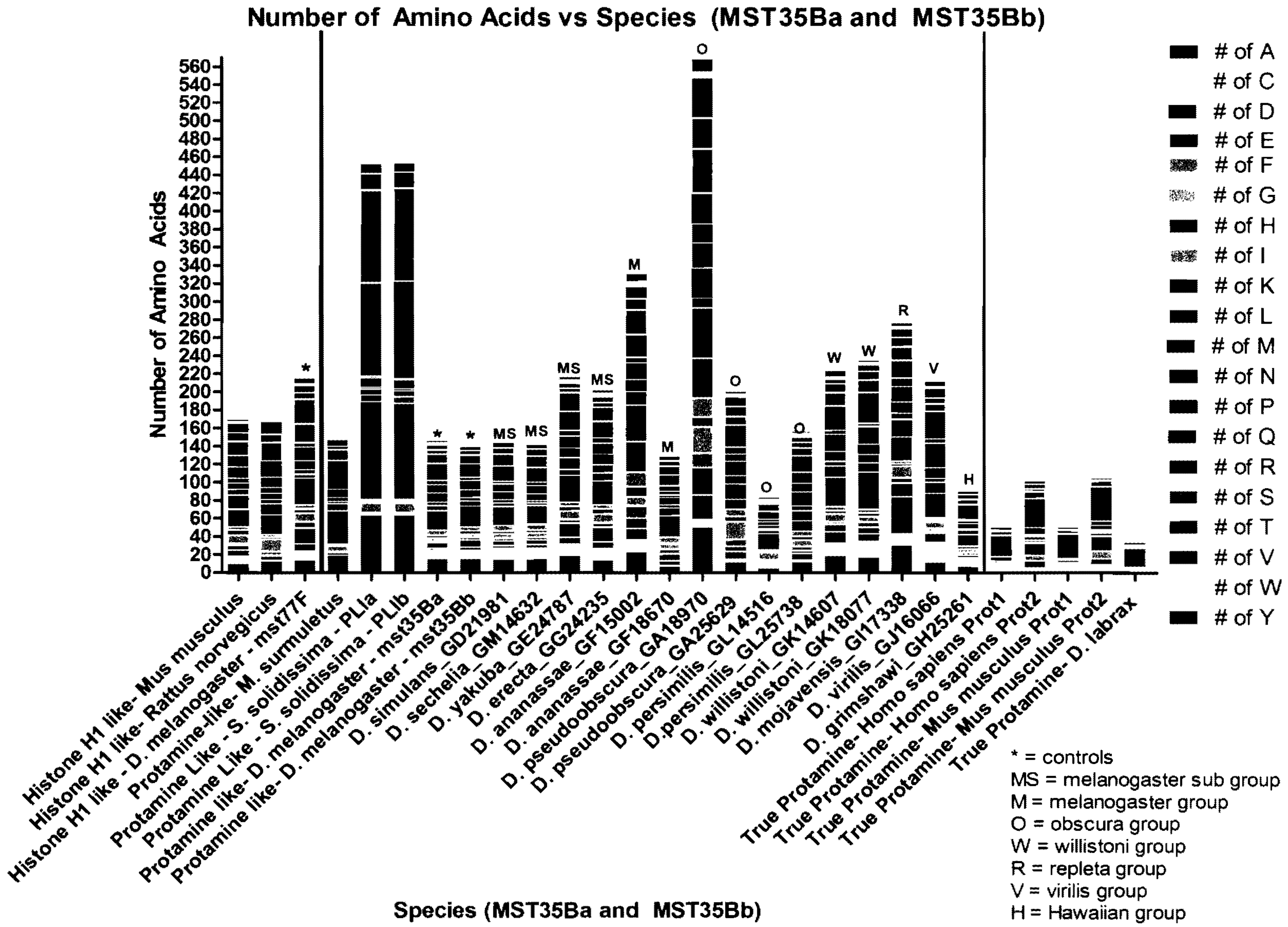


Figure 12A. Amino Acid percentage versus Species (Mst35Ba conserved region)

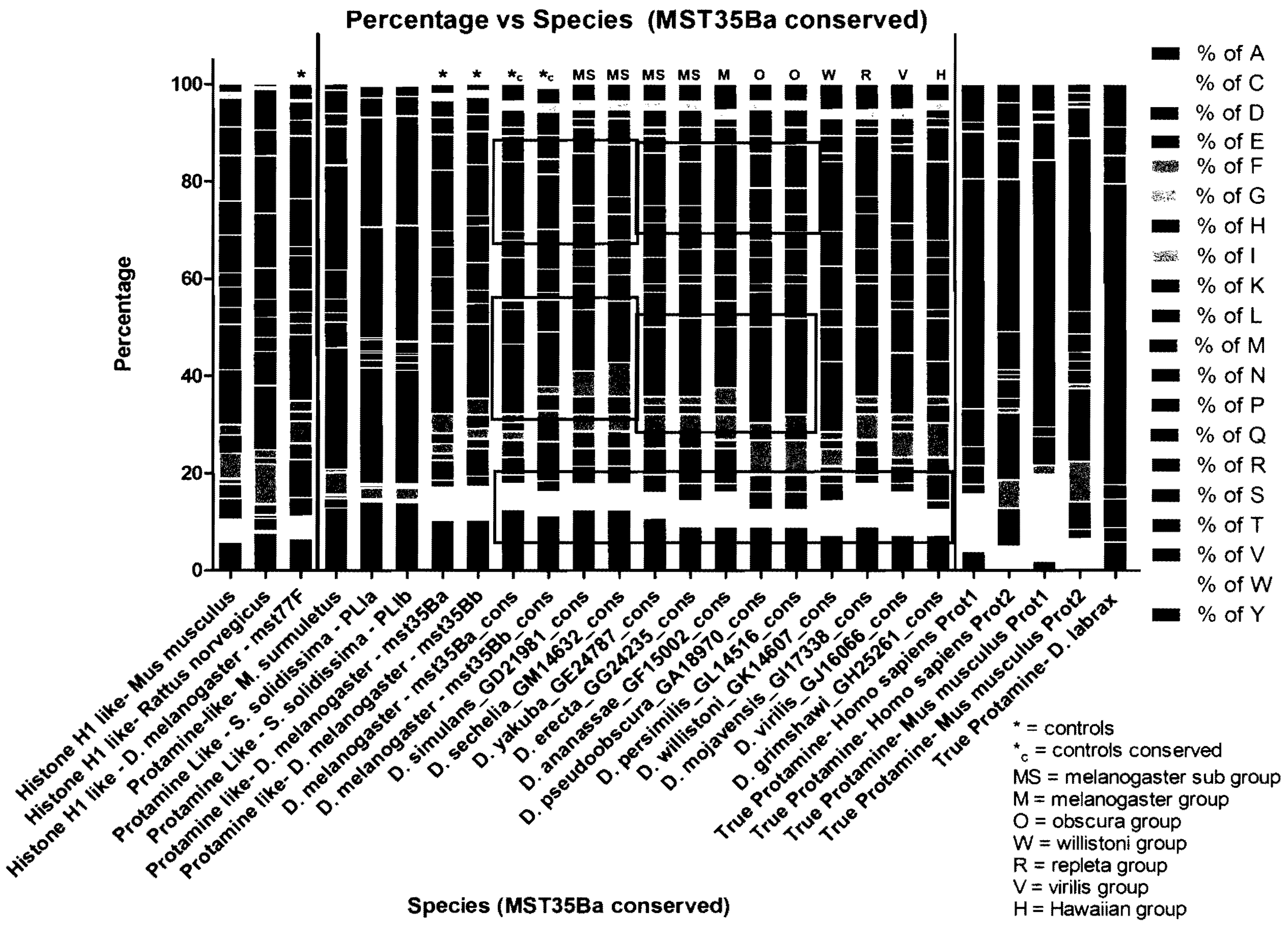


Table 1B. Amino Acid percentage versus Species (Mst35Ba conserved region)

Species	Percentage of Amino Acids													
	% A	**%C**	**% E**	**% G**	**% H**	**% K**	**% L**	**% N**	**% P**	**% Q**	**% R**	**% S**	**% T**	**% V**
Histone H1 like- Mus musculus	5.88	4.71	2.94	5.29	3.53	11.18	9.41	4.12	2.94	7.65	7.06	9.41	5.88	5.88
Histone H1 like- Rattus norvegicus	7.69	0.59	0.59	8.28	1.18	13.02	7.10	4.14	3.55	6.51	11.24	11.83	5.33	8.28
Histone H1 like - D. melanogaster - mst77F	6.51	4.65	7.91	4.65	1.86	13.49	2.33	4.65	6.98	1.86	9.77	13.02	3.26	3.72
Protamine-like- M. surmuletus	12.75	0.00	0.00	4.70	0.00	24.83	5.37	2.68	6.04	0.00	21.48	8.05	2.68	4.70
Protamine-like - S. solidissima - PLla	14.10	0.22	0.00	2.42	0.44	23.79	1.54	0.44	2.20	0.44	22.69	22.69	3.96	2.42
Protamine-like - S. solidissima - PLlb	14.07	0.22	0.00	2.42	0.44	23.30	1.54	0.44	2.20	0.22	23.74	22.64	3.96	2.20
Protamine-like- D. melanogaster - Mst35Ba	10.27	6.85	4.11	2.05	2.05	14.38	4.11	6.85	6.16	3.42	12.33	7.53	3.42	3.42
Protamine-like- D. melanogaster - Mst35Bb	10.42	6.94	5.56	2.08	2.78	15.28	4.17	5.56	7.64	2.08	10.42	6.94	3.47	3.47
D. melanogaster - Mst35Ba_cons	12.50	5.36	3.57	1.79	1.79	14.29	7.14	8.93	3.57	1.79	14.29	5.36	1.79	3.57
D. melanogaster - Mst35Bb_cons	11.29	4.84	8.06	1.61	3.32	11.29	6.45	4.84	4.84	3.23	11.29	3.23	4.84	4.84
D. simulans_GD21981_cons	12.50	5.36	3.57	3.57	3.57	12.50	5.36	3.57	5.36	3.57	10.71	5.36	1.79	1.79
D. sechelia_GM14632_cons	12.50	5.36	3.57	3.57	3.57	12.50	5.36	3.57	5.36	3.57	10.71	5.36	1.79	0.00
D. yakuba_GE24787_cons	10.71	5.36	5.36	3.57	1.79	14.29	7.14	5.36	7.14	3.57	10.71	3.57	1.79	3.57
D. erecta_GG24235_cons	8.93	5.36	7.14	3.57	1.79	16.07	7.14	5.36	5.36	3.57	8.93	3.57	3.57	3.57
D. ananassae_GF15002_cons	8.93	7.14	7.14	5.36	1.79	12.50	5.36	5.36	5.36	0.00	16.07	3.57	0.00	1.79
D.pseudoobscura_GA18970_cons	8.93	3.57	3.57	7.14	1.79	19.64	7.14	5.36	3.57	3.57	7.14	7.14	3.57	5.36
D. persimilis_GL14516_cons	8.93	3.57	3.57	7.14	1.79	19.64	7.14	5.36	3.57	3.57	5.36	8.93	3.57	3.57
D. willistoni_GK14607_cons	7.14	7.14	1.79	3.57	1.79	14.29	7.14	8.93	7.14	0.00	14.29	1.79	3.57	3.57
D. mojavensis_GI17338_cons	8.93	8.93	3.57	5.36	1.79	14.29	8.93	5.36	7.14	3.57	12.50	0.00	1.79	1.79
D. virilis_GJ16066_cons	7.14	8.93	3.57	5.36	1.79	12.50	8.93	5.36	7.14	3.57	14.29	1.79	1.79	3.57
D. grimshawi_GH25261_cons	7.14	5.36	5.36	7.14	3.57	7.14	8.93	7.14	3.57	3.57	16.07	7.14	1.79	1.79
True Protamine- Homo sapiens Prot1	3.92	11.76	0.00	0.00	1.96	0.00	0.00	0.00	3.92	7.84	47.06	9.80	1.96	0.00
True Protamine- Homo sapiens Prot2	0.00	4.90	7.84	5.88	13.73	1.96	3.92	0.00	0.98	7.84	31.37	7.84	2.94	4.90
True Protamine- Mus musculus Prot1	1.96	17.65	0.00	0.00	0.00	5.88	0.00	0.00	0.00	0.00	54.90	7.84	1.96	0.00
True Protamine- Mus musculus Prot2	0.00	6.54	5.61	8.41	14.95	2.80	1.87	0.00	3.74	4.67	35.51	6.54	0.93	1.87
True Protamine- D. labrax	5.88	0.00	2.94	0.00	0.00	0.00	0.00	0.00	5.88	2.94	61.76	5.88	5.88	8.82

Figure 12B. Amino Acid percentage versus Species (Mst35Bb conserved region)

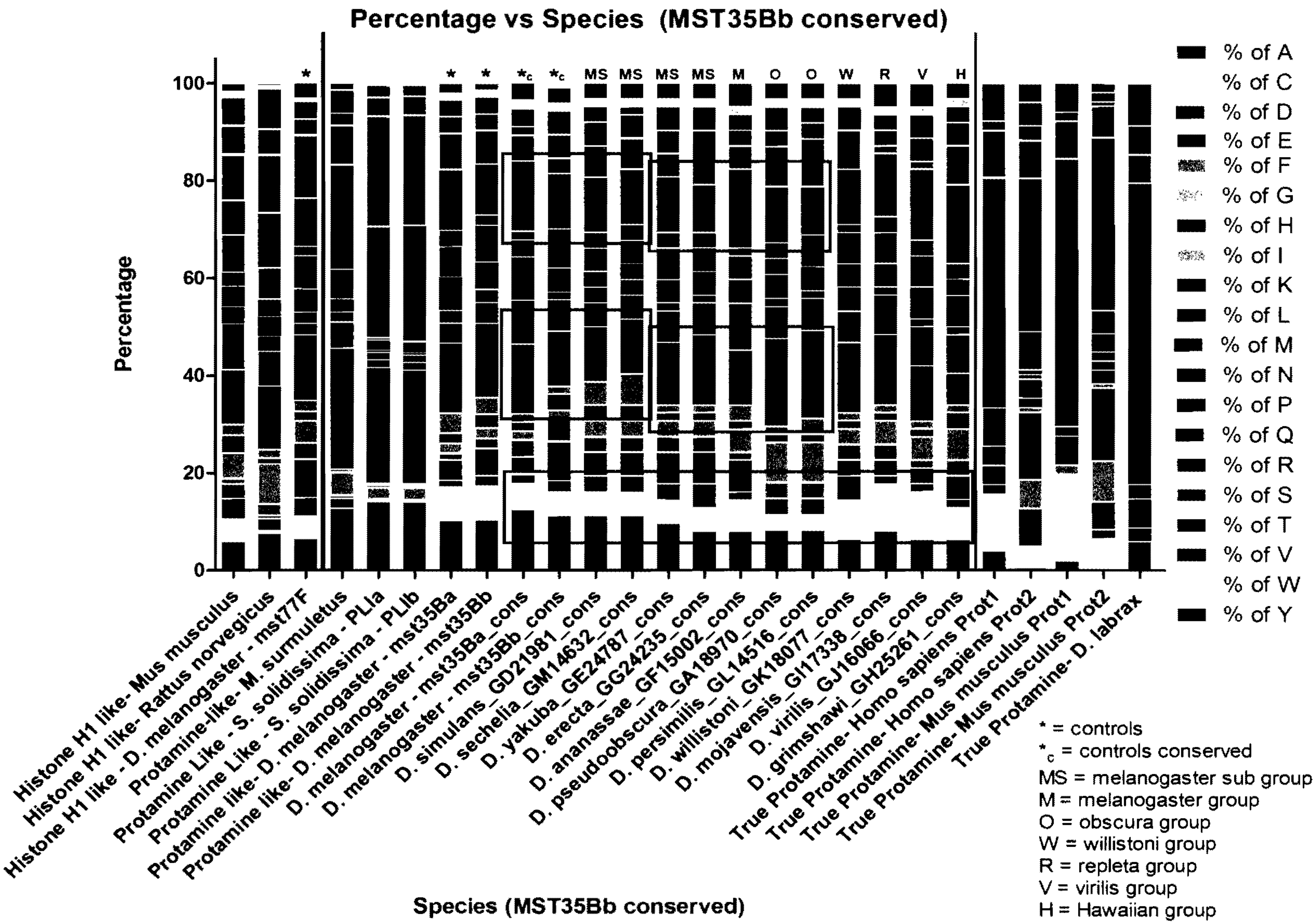


Table 1C. Amino Acid percentage versus Species (Mst35Bb conserved region)

Species	Percentage of Amino Acids													
	% A	**% C**	**% E**	**% G**	**% H**	**% K**	**% L**	**% N**	**% P**	**% Q**	**% R**	**% S**	**% T**	**% V**
Histone H1 like- Mus musculus	5.88	4.71	2.94	5.29	3.53	11.18	9.41	4.12	2.94	7.65	7.06	9.41	5.88	5.88
Histone H1 like- Rattus norvegicus	7.69	0.59	0.59	8.28	1.18	13.02	7.10	4.14	3.55	6.51	11.24	11.83	5.33	8.28
Histone H1 like - D. melanogaster - mst77F	6.51	4.65	7.91	4.65	1.86	13.49	2.33	4.65	6.98	1.86	9.77	13.02	3.26	3.72
Protamine-like- M. surmuletus	12.75	0.00	0.00	4.70	0.00	24.83	5.37	2.68	6.04	0.00	21.48	8.05	2.68	4.70
Protamine-like - S. solidissima - PLla	14.10	0.22	0.00	2.42	0.44	23.79	1.54	0.44	2.20	0.44	22.69	22.69	3.96	2.42
Protamine-like - S. solidissima - PLlb	14.07	0.22	0.00	2.42	0.44	23.30	1.54	0.44	2.20	0.22	23.74	22.64	3.96	2.20
Protamine-like- D. melanogaster - Mst35Ba	10.27	6.85	4.11	2.05	2.05	14.38	4.11	6.85	6.16	3.42	12.33	7.53	3.42	3.42
Protamine-like- D. melanogaster - Mst35Bb	10.42	6.94	5.56	2.08	2.78	15.28	4.17	5.56	7.64	2.08	10.42	6.94	3.47	3.47
D. melanogaster - Mst35Ba_cons	12.50	5.36	3.57	1.79	1.79	14.29	7.14	8.93	3.57	1.79	14.29	5.36	1.79	3.57
D. melanogaster - Mst35Bb_cons	11.29	4.84	8.06	1.61	3.32	11.29	6.45	4.84	4.84	3.23	11.29	3.23	4.84	4.84
D. simulans_GD21981_cons	11.29	4.84	4.84	3.23	3.23	11.29	4.84	3.23	4.84	3.23	11.29	6.45	4.84	3.23
D. sechelia_GM14632_cons	11.28	4.84	4.84	3.23	3.23	11.29	4.84	3.23	4.84	3.23	11.29	6.45	4.84	1.61
D. yakuba_GE24787_cons	9.68	4.84	6.45	3.23	1.61	12.90	6.45	4.84	6.45	3.23	11.29	4.84	4.84	4.84
D. erecta_GG24235_cons	8.06	4.84	6.45	3.23	1.61	14.52	6.45	4.84	4.84	3.23	9.68	4.84	6.45	4.84
D. ananassae_GF15002_cons	8.06	6.45	6.45	4.84	1.61	11.29	4.84	4.84	4.84	1.61	16.13	4.84	3.23	3.23
D. pseudoobscura_GA18970_cons	8.20	3.28	3.28	8.20	1.64	18.03	6.56	4.92	3.28	3.28	11.48	8.20	3.28	4.92
D. persimilis_GL14516_cons	8.20	3.28	3.28	8.20	1.64	18.03	6.56	4.92	3.28	3.28	9.84	9.84	3.28	3.28
D. willistoni_GK18077_cons	6.45	8.06	3.23	3.23	1.61	14.52	6.45	4.84	6.45	1.61	11.29	8.06	0.00	4.84
D. mojavensis_GI17338_cons	8.06	9.68	3.23	4.84	1.61	14.52	8.06	4.84	6.45	3.23	12.90	1.61	3.23	3.23
D. virilis_GJ16066_cons	6.45	9.68	3.23	4.84	1.61	11.29	8.06	6.45	6.45	3.23	14.52	1.61	4.84	4.84
D. grimshawi_GH25261_cons	6.45	6.45	4.84	6.45	3.23	6.45	8.06	6.45	3.23	3.23	16.13	8.06	4.84	3.23
True Protamine- Homo sapiens Prot1	3.92	11.76	0.00	0.00	1.96	0.00	0.00	0.00	3.92	7.84	47.06	9.80	1.96	0.00
True Protamine- Homo sapiens Prot2	0.00	4.90	7.84	5.88	13.73	1.96	3.92	0.00	0.98	7.84	31.37	7.84	2.94	4.90
True Protamine- Mus musculus Prot1	1.96	17.65	0.00	0.00	0.00	5.88	0.00	0.00	0.00	0.00	54.90	7.84	1.96	0.00
True Protamine- Mus musculus Prot2	0.00	6.54	5.61	8.41	14.95	2.80	1.87	0.00	3.74	4.67	35.51	6.54	0.93	1.87
True Protamine- D. labrax	5.88	0.00	2.94	0.00	0.00	0.00	0.00	0.00	5.88	2.94	61.76	5.88	5.88	8.82

III. Functional analysis of the conserved region

Using three different bioinformatics tools, functional analyses of the conserved regions found in Mst35Ba and Mst35Bb for all 12 Drosophila species were carried out. Additionally, an alignment of all of the conserved matches' secondary structures to their respective conserved regions (Mst35Ba conserved and Mst35Bb conserved) was created through the use of Molsoft ICM-Browser.

Our results indicate that the conserved region found in Mst35Ba and Mst35Bb is a DNA binding domain. Analysis using DNA-Binder indicated that the majority of the putative protamine-like protein sequences contained a DNA-binding domain or had a chance to be a DNA binding protein except for *D. ananassa_GF15002*, which at most a slim chance of being a DNA-binding protein. *D. pseudoobscura GA18070* had a small chance of being a DNA-binding protein, which may be attributed to the low coverage score with the genomic sequence. All conserved regions for each match and the controls had high likelihood of being a DNA-Binding domain (Table 2).

Using BindN+ and BindN-RF (Random Forests) were able to predict the actual residues with a score to be DNA binding or non-DNA binding. BindN+ is able to support vector machines (algorithm) for its prediction.

Figures 13A and 13B, illustrate the conserved region (shaded region) for Mst35Ba and Mst35Bb for BindN-RF (strict) and BindN+ (relaxed). The conserved region corresponds directly to conserved region that contains many putative DNA-binding residues for the *D. pseudoobscura* matches shown in Figures 14A and 14B. As indicated in Figures 14A and 14B, the conserved regions are predicted to have majority of its residues be DNA binding for the two *D. pseudoobscura* best matches. In Table 3, the majority of the matches have residues that have been predicted to be DNA binding, except for the best match for *D. ananassae, D.*

pseudoobscura, and D. mojavensis. These matches' DNA binding residue percentages range from 15% to 47%. The low DNA binding residue percentage for these three matches can be attributed to their large number of amino acid present in their respective protein. Hence the increase in variability is main reason for their low DNA binding residue percentage for these three matches (Figs. 11A, 11B, and Table 1A). Additionally, the conserved regions shown in Figures 14A and 14B indicate that the majority of the putative DNA binding residues belong to the conserved region. The conserved regions for each matching sequence have a large percentage of basic amino acids that are more likely involved in DNA binding.

In Figure 15, through Swiss Model Interpro Domain Scan, the functional groups for Mst35Ba, Mst35Bb, *D. pseudoobscura* matches and their conserved regions were graphically visualized. This illustrates that *D. pseudoobscura* matches contain a high mobility group that overlaps with the high mobility group box. Furthermore, the presence of the high mobility group is present for every matched species. As *D. pseudoobscura* is our experimental fly, only this data has been shown. Overall this indicates that the high mobility group box has been present in many DNA binding proteins and regions.

Using Protein Homology/analogy Recognition Engine 2.0 (Phyre 2) further analysis on the protein matches and the conserved regions were conducted. Phyre 2takes a protein sequences and predicts its 3D structure. The protein sequence is searched through a database containing 10 million known sequences for homologues through the use of PSI-BLAST to examine the evolutionary relationship with the known sequences.

In Table 4A illustrates a sample of highly detailed analysis conserved matches for Mst35Ba. All three sample matches (c2e6oA, c2cs1A, d1v64a) overlap through protein of unknown function DUF1074 Family and the high mobility group box. Likewise in Table 4B, which illustrates analysis of the whole protein a few sample results show the same occurrence of DUF1074 of protein of unknown function family overlapping with the HMG box. There is a possibility that there is some relationship occurring here. (Fig. 15)

Using Phyre 2, all of the 3D secondary wire frame structures of the conserved regions were saved and analyzed through Molsoft ICM Browser. This analysis curtailed the alignment of all respective matches on top of each based on their respective conserved regions of the controls (Mst35Ba and Mst35Bb). Figure 16 shows that all the conserved regions among the 12 sequenced species have similar tertiary alignment of the three alpha helices. The conserved region occurred in high confidence region for majority of the matches (not shown). Additionally, the conserved region is very similar to known functional groups that have been presented in Figure 18. All of these functional groups are HMG boxes and are involved transcription or DNA-binding. Lastly these functional groups are found in the conserved region matches among the 12 species as well.

Table 2. DNA-Binder Predictions for Mst35Ba, Mst35Bb, and conserved sequences

Sequence Name	Realistic Dataset[1]		Alternative Dataset[2]		Main Dataset[3]	
SVM threshold = -1	SVM Score	DNA Bind (Y/N/M)	SVM Score	DNA Bind (Y/N/M)	SVM Score	DNA (Y/N)
Dmel_Mst35Ba	2.5352528	Yes	0.4838234	Yes	-	-
D.sim_GD21981	1.5765025	Yes	0.55971467	Maybe/Yes	-	-
D.sec_GM14632	0.97788236	Yes	0.59466325	Maybe/Yes	-	-
D.yak_GE24787	-1.0898525	No	0.26950508	Maybe	-	-
D.ere_GG24235	-0.55590193	Maybe/No	0.53697442	Maybe/Yes	-	-
D.ana_GF15002	-0.75164668	No	-0.27363754	Maybe	-	-
D.pse_GA18970	-0.81001403	No	0.57353094	Maybe/Yes	-	-
D.pse_GA25629	0.34727124	Maybe	0.19297556	Maybe	-	-
D.per_GL14516	2.0164925	Yes	0.93872436	Yes	-	-
D.will_GK14607	0.73459339	Yes	1.0851398	Yes	-	-
D.moj_GI17338	-1.4897	No	0.77375851	Yes	-	-
D.vir_GJ16066	1.9136553	Yes	0.070493013	Maybe	-	-
D.gri_GH25261	3.0519743	Yes	0.070493013	Maybe	-	-
D.will_GK18077	2.6063606	Yes	0.13364001	Maybe	-	-
D.mel_Mst35Bb	1.5465344	Yes	0.80853348	Yes	-	-
Dmel_Mst35Ba_cons	4.6082969	Yes	-	-	3.2052009	Yes
D.sim_GD21981_ba_C	1.6854101	Yes	-	-	2.0974391	Yes
D.sec_GM14632_ba_C	2.1782354	Yes	-	-	2.2391386	Yes
D.yak_GE24787_ba_C	2.501675	Yes	-	-	2.1575922	Yes
D.ere_GG24235_ba_C	2.239147	Yes	-	-	2.1354235	Yes
D.ana_GF15002_ba_C	1.9012743	Yes	-	-	2.7949326	Yes
D.pse_GA18970_ba_C	2.3441848	Yes	-	-	1.8246887	Yes
Dpse_GA25629_ba_C	2.5587343	Yes	-	-	2.0852804	Yes
D.per_GL14516_ba_C	1.2632506	Yes	-	-	1.5922416	Yes
D.will_GK14607_ba_C	4.2431542	Yes	-	-	2.5043752	Yes
D.moj_GI17338_ba_C	4.015309	Yes	-	-	2.5289927	Yes
D.vir_GJ16066_ba_C	3.58217	Yes	-	-	2.7286503	Yes
D.gri_GH25261_ba_C	2.0370817	Yes	-	-	2.8598303	Yes
D.mel_Mst35Bb_cons	1.9513811	Yes	-	-	2.5348116	Yes
D.will_GK18077_bb_C	1.7061587	Yes	-	-	2.1913375	Yes
D.sim_GD21981_bb_C	1.3290914	Yes	-	-	2.1786021	Yes
D.sec_GM14632_bb_C	1.661014	Yes	-	-	2.3063786	Yes
D.yak_GE24787_bb_C	1.8465637	Yes	-	-	2.2325163	Yes
D.ere_GG24235_bb_C	1.8013479	Yes	-	-	2.0569327	Yes
D.ana_GF15002_bb_C	1.5110199	Yes	-	-	2.8097756	Yes
D.pse_GA18970_bb_C	2.9576456	Yes	-	-	2.4946569	Yes
Dpse_GA25629_bb_C	2.5587343	Yes	-	-	2.0852804	Yes
D.per_GL14516_bb_C	2.2245383	Yes	-	-	2.2800249	Yes
D.moj_GI17338_bb_C	3.7578127	Yes	-	-	2.6700481	Yes
D.vir_GJ16066_bb_C	3.4336474	Yes	-	-	2.5675003	Yes
D.gri_GH25261_bb_C	1.6257282	Yes	-	-	2.7873782	Yes

Figure 13A.	
DNA Binding Residues according to Bind N+ for Mst35Ba (Prot A)	
MSSNNVNECKSLWNGIISISAKDESPKGLTEMCNHPIRRAPQKCKPMKSCAKPRRKAACAKATRPKVKCAPR -++++-++-++-++---+-+-+--+++--+----++++++++-+++++++++++++++-++++++-++-++ 269742223252343345333722236234345323279326334335533869894343826935232228 QK[redacted]VTTSERHKRRRICQ ++-++++++++++++ 5536887988989323	
DNA Binding Residues according to BindN-RF for Mst35Ba (Prot A)	
MSSNNVNECKSLWNGIISISAKDESPKGLTEMCNHPIRRAPQKCKPMKSCAKPRRKAACAKATRPKVKCAPRQ -++++-+--++-++---+-+++-+++++-+----+--++--++-+++++-+++++++--++++++-+-+-++ 4675474367633352773545225252443982245792438383296239599923249359494923275 K[redacted]VTTSERHKRRRICQQY +-++++++++++-++- 84646397887932232	
Figure 13B.	
DNA Binding Residues according to Bind N+ for Mst35Bb (Prot B)	
MSSNNVNECKSLWNGIISISAKDESPKGLTEMCNHPKRRAPPKCKPMKSCAKPRRKAACAKATRPKVKCAPS -++++-++-++-++---+-+-+--+++--+----++++++++-+++++++++++++++-++++++-++-++ 269742223252343345333722236234345333578325334335533869894343826935232236 QK[redacted]HKRRRICK ++++++++++ 5688999224	
DNA Binding Residues according to BindN-RF for Mst35Bb (Prot B)	
MSSNNVNECKSLWNGIISISAKDESPKGLTEMCNHPKRRAPPKCKPMKSCAKPRRKAACAKATRPKVKCAPS -++++-+--++-++---+-+++-+++++-+----+--++---+-+++++-+++++++--++++++-+-+-+ 467547436763335277354522525244399224379232838329623959992324935949393227 QK[redacted]HKRRRICK ++++++++++ 5878879326	
Key:	Red and + indicates DNA Binding Residues Green and - indicates a non-DNA Binding regions Dark Shaded region is the conserved region
Confidence Score:	0 (lowest) to 9 (highest)

Figure 14A.

DNA Binding Residues based on Bind N+ for *D. pseudoobscura GA18970*

```
MGCSSKPQKKYR[redacted]GKKSTAMDILIREEDISLAQIGVYASVSFLVVSAVGAALYTTCSRRY
--+++++++++             +++----------------------+-------------------+-
232279557657            35233889999376562453586252668989578887876688617
RLNWFEQNLLESANEKDEDQQREALVAGAAGYNVDNLNECSRGNLSPTSLKNDENDPAFWVPASVASTAAIQQQVSNTTEESAPPTPTSPTGSLKSNTLSLCSTASVPIARSDKHVVLAM
-----------+----------------------------------------------------------------------------------+++--+--------------------
588688543342737689989899898465328986987852937784496534446788989786658483329464757598728435344423223349648799594799999999
HPTRPRVSSMNAKLDHTKIDMTLYRSHAQPKTLDPAPAIEVRGNLHVGISYDPVGGLLNVRLLEAQNLQPRQFSGSADPYAKVRLLPDKKNFWQTRIHKKTLNPVFDEHFVFEVAAGVIC
--------------------------------------------------------------------------+-----------------++++------------------------
544362753669686987999796477779999999999895947587987879376839488773575636323224529584999854324234222225536888899999998977
KRTVEILLYDFDAYSRHVCIGGTKLHLANIDLSEQLQLWTPLSSASAQDMKVDLGDIMVSLAYLPSAERLMVVLIKARNLRIVDDARNSSDPYVKVTLLGPVGKKMKKRKTGVQRSTVNP
--------------------------------------------------------------------------------++----++++------------+--++--+-+++++---+
4879898859672433278888968889879978795786877577889989899999895872685987999989892253222324332243948496857433623224222422323
VYNEALAFDVNKETLKNCVLEFTVVHDGLLGSSEILGRTLIGNSSEVRTEEKIFFEEMFRAKNATAQWVPLQEPATNLANAAKSTTNKN
-----------------------------+-+++------------------------------------------------+--++++
33368899996778887999896857572222447886776674539588979999999374553655977768755666782422232
```

DNA Binding Residues based on BindN-RF for *D. pseudoobscura GA18970*

```
MGCSSKPQKKYR[redacted]GKKSTAMDILIREEDISLAQIGVYASVSFLVVSAVGAALYTTCSRRY
-+--++++++++            +++----------------------+-------------------+-
523259343856            35233889999376562453586252668989578887876688617
RLNWFEQNLLESANEKDEDQQREALVAGAAGYNVDNLNECSRGNLSPTSLKNDENDPAFWVPASVASTAAIQQQVSNTTEESAPPTPTSPTGSLKSNTLSLCSTASVPIARSDKHVVLAM
-----------+----------------------------------------------------------------------------------+++--+--------------------
588688543342737689989899898465328986987852937784496534446788989786658483329464757598728435344423223349648799594799999999
HPTRPRVSSMNAKLDHTKIDMTLYRSHAQPKTLDPAPAIEVRGNLHVGISYDPVGGLLNVRLLEAQNLQPRQFSGSADPYAKVRLLPDKKNFWQTRIHKKTLNPVFDEHFVFEVAAGVIC
--------------------------------------------------------------------------+-----------------++++------------------------
544362753669686987999796477779999999999895947587987879376839488773575636323224529584999854324234222225536888899999998977
KRTVEILLYDFDAYSRHVCIGGTKLHLANIDLSEQLQLWTPLSSASAQDMKVDLGDIMVSLAYLPSAERLMVVLIKARNLRIVDDARNSSDPYVKVTLLGPVGKKMKKRKTGVQRSTVNP
--------------------------------------------------------------------------------++----++++------------+--++--+-+++++---+
4879898859672433278888968889879978795786877577889989899999895872685987999989892253222324332243948496857433623224222422323
VYNEALAFDVNKETLKNCVLEFTVVHDGLLGSSEILGRTLIGNSSEVRTEEKIFFEEMFRAKNATAQWVPLQEPATNLANAAKSTTNKN
-----------------------------+-+++------------------------------------------------+--++++
33368899996778887999896857572222447886776674539588979999999374553655977768755666782422232
```

Figure 14B.

DNA Binding Residues based on Bind N+ and Bind-RF for *D. pseudoobscura GA25629*

```
MAPVMKLRNPFLNFLDVYRRNHSNMNMVTAAKAGAQRWRHLTDEQRSKFRRNVDMDCHGSGLDSRKRKRG[redacted]
-----+-++--------+++++++++-++-+++++++++-+++++++++++-------++++++++++++
7954452742373356224988774323422623368782432225595964255432334638986965
[redacted]SGPRYIPVRVSIEMNTQEILFTGLQTGHETSNCKEALINGGGGGGGGKAMPTIRLFHLVCFNSTMVRTLWQSGRREM
+++++-+-+---++-------+--+++-+++-+----+++++++++++-++-+--------++++-++-+++++++-
72454426573235223588822324452446343544447787954932486244687342342444233628832
```

DNA Binding Residues based on Bind-RF for *D. pseudoobscura GA25629*

```
MAPVMKLRNPFLNFLDVYRRNHSNMNMVTAAKAGAQRWRHLTDEQRSKFRRNVDMDCHGSGLDSRKRKRG[redacted]
-----+-++---+----+++++++-+--+--++++++++++++--+++++++-----+-+++-++++++
8922494862664877437676654454534824349583343627583886252442343326975864
[redacted]SGPRYIPVRVSIEMNTQEILFTGLQTGHETSNCKEALINGGGGGGGGKAMPTIRLFHLVCFNSTMVRTLWQSGRREM
+++++---+-+---++-----+---+++-+++-+----+---++++-+--++-+-------+++--++-+++++++-
72263323835336233598923454253455543576422233523834255777349463563784544738934
```

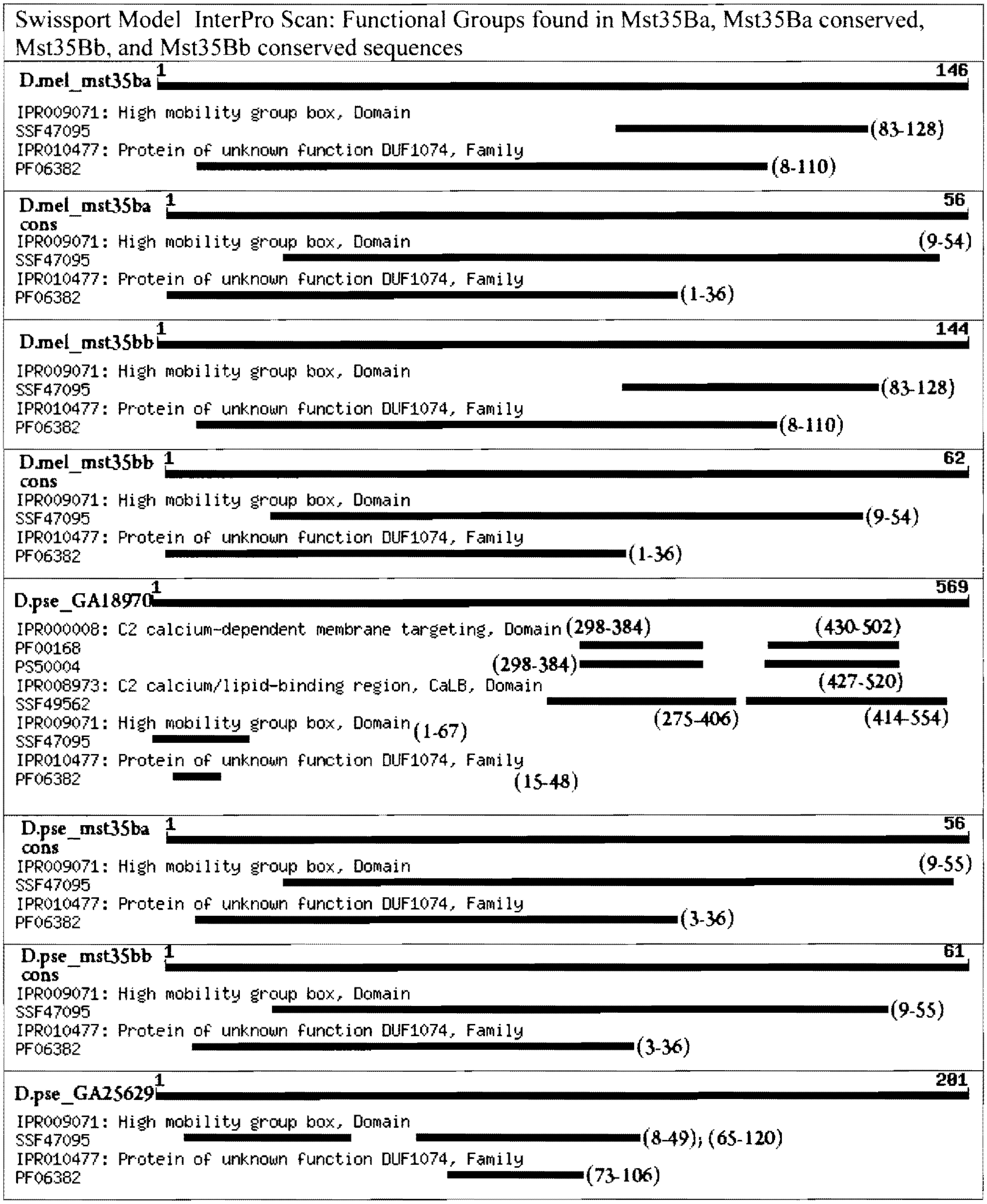
Key:	Red and + indicates DNA Binding Residues Green and - indicates a non-DNA Binding regions Dark Shaded Region is the conserved region Blue Shaded Region is the extended conserved region for only Mst35Bb
Confidence Score:	0 (lowest) to 9 (highest)

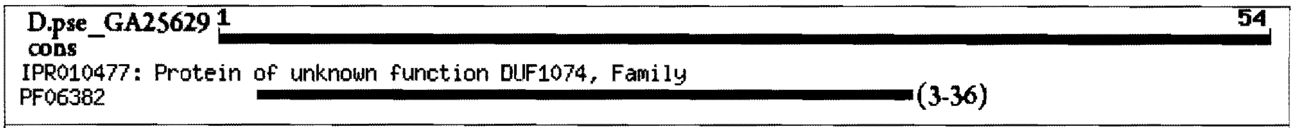
Table 3. DNA Binding Residues based on BindN+ and Bind-RF for Mst35Ba and Mst35Bb

Sequence	BindN+			BindN-RF		
	Specificity set to recommended 79% Estimated Sensitivity was 80.28%			specificity set to recommended 78.22% Estimated Sensitivity was 78.03		
	Sequence Length (Amino Acids)	Predicted Binding Site (Amino Acids)	Percentage of DNA Binding Sites	Sequence Length (Amino Acids)	Predicted Binding Site (Amino Acids)	Percentage of DNA Binding Sites
Dmel_Mst35Ba	146	106	72.60273973	146	91	62.32876712
D.sim_GD21981	147	102	69.3877551	146	88	60.2739726
D.sec_GM14632	147	101	68.70748299	147	90	61.2244898
D.yak_GE24787	217	125	57.60368664	217	118	54.37788018
D.ere_GG24235	202	115	56.93069307	202	112	55.44554455
D.ana_GF15002	331	155	46.82779456	331	144	43.50453172
D.pse_GA18970	569	88	15.46572935	569	90	15.8172232
D.pse_GA25629	201	122	60.69651741	201	106	52.73631841
D.per_GL14516	84	57	67.85714286	84	57	67.85714286
D.will_GK14607	224	153	68.30357143	224	144	64.28571429
D.moj_GI17338	277	85	30.68592058	277	89	32.1299639
D.vir_GJ16066	212	127	59.90566038	212	133	62.73584906
D.gri_GH25261	93	66	70.96774194	93	61	65.59139785
D.will_GK18077	235	170	72.34042553	235	155	65.95744681
D.mel_Mst35Bb	144	100	69.44444444	144	88	61.11111111

As table 3, both BindN+ and BindN-RF reveal similar percentage for the DNA Binding Sites. Overall there the difference is approximately the range of difference is between 0 to 13% between BindN+ and BindN-RF. BindN+ algorithm is based upon support vector machines and is a more relaxed algorithm when compared to stricter algorithm of BindN-RF (Random Forests).

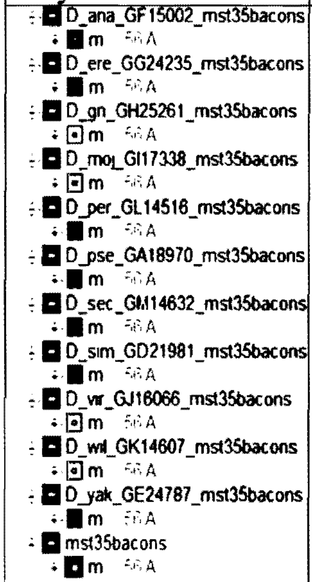
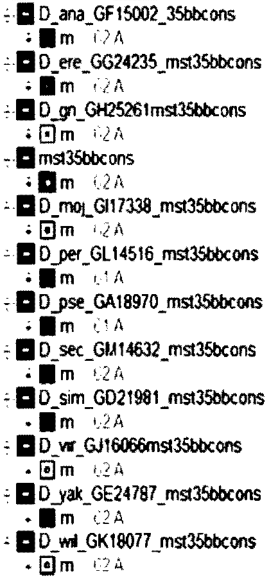
Figure 15. Functional Groups found in Mst35Ba, Mst35Ba conserved, Mst35Bb, and Mst35Bb conserved


Swissport Model InterPro Scan: Functional Groups found in Mst35Ba, Mst35Ba conserved, Mst35Bb, and Mst35Bb conserved sequences
D.mel_mst35ba 1 146
IPR009071: High mobility group box, Domain
SSF47095 (83-128)
IPR010477: Protein of unknown function DUF1074, Family
PF06382 (8-110)
D.mel_mst35ba cons 1 56
IPR009071: High mobility group box, Domain (9-54)
SSF47095
IPR010477: Protein of unknown function DUF1074, Family
PF06382 (1-36)
D.mel_mst35bb 1 144
IPR009071: High mobility group box, Domain
SSF47095 (83-128)
IPR010477: Protein of unknown function DUF1074, Family
PF06382 (8-110)
D.mel_mst35bb cons 1 62
IPR009071: High mobility group box, Domain
SSF47095 (9-54)
IPR010477: Protein of unknown function DUF1074, Family
PF06382 (1-36)
D.pse_GA18970 1 569
IPR000008: C2 calcium-dependent membrane targeting, Domain (298-384) (430-502)
PF00168
PS50004 (298-384)
IPR008973: C2 calcium/lipid-binding region, CaLB, Domain (427-520)
SSF49562
IPR009071: High mobility group box, Domain (1-67) (275-406) (414-554)
SSF47095
IPR010477: Protein of unknown function DUF1074, Family
PF06382 (15-48)
D.pse_mst35ba cons 1 56
IPR009071: High mobility group box, Domain (9-55)
SSF47095
IPR010477: Protein of unknown function DUF1074, Family
PF06382 (3-36)
D.pse_mst35bb cons 1 61
IPR009071: High mobility group box, Domain
SSF47095 (9-55)
IPR010477: Protein of unknown function DUF1074, Family
PF06382 (3-36)
D.pse_GA25629 1 281
IPR009071: High mobility group box, Domain
SSF47095 (8-49); (65-120)
IPR010477: Protein of unknown function DUF1074, Family
PF06382 (73-106)



All matches for Mst35Ba (Prot A) and Mst35Bb(Prot B) contained HMG and DUF1074 (a protein of unknown function). There is an overlap of these functional groups in their respective matches. Thus, there is a possibility that there HMG group and DUF1074 could be involved in some DNA condensating process.

Figurer 16. Conservation of Secondary Structures among all Mst35Ba and Mst35Bb

	All Matches for:
Key:	Mst35Ba (Prot A)
 D_ana_GF15002_mst35bacons m 56 A D_ere_GG24235_mst35bacons m 56 A D_gn_GH25261_mst35bacons m 56 A D_moj_GI17338_mst35bacons m 56 A D_per_GL14516_mst35bacons m 56 A D_pse_GA18970_mst35bacons m 56 A D_sec_GM14632_mst35bacons m 56 A D_sim_GD21981_mst35bacons m 56 A D_vir_GJ16066_mst35bacons m 56 A D_wil_GK14607_mst35bacons m 56 A D_yak_GE24787_mst35bacons m 56 A mst35bacons m 56 A	
Key:	Mst35Bb (Prot B)
 D_ana_GF15002_35bbcons m 62 A D_ere_GG24235_mst35bbcons m 62 A D_gn_GH25261mst35bbcons m 62 A mst35bbcons m 62 A D_moj_GI17338_mst35bbcons m 62 A D_per_GL14516_mst35bbcons m 61 A D_pse_GA18970_mst35bbcons m 61 A D_sec_GM14632_mst35bcons m 62 A D_sim_GD21981_mst35bbcons m 62 A D_vir_GJ16066mst35bbcons m 62 A D_yak_GE24787_mst35bbcons m 62 A D_wil_GK18077_mst35bbcons m 62 A	

Wireframe secondary structure of all respective SNBP matches overlaid on each other illustrating the conservation of secondary and tertiary structure within the conserved region among different Drosophila species.

Table 4A. Detailed Analysis of Functional Groups found in Mst35Ba conserved matches

Sample Matches for Mst35Ba_cons		*Mst35 Ba cons*	*Mst35 Bb cons*	*D.sim con. GD21981*	*D.sec con. GM14632*	*D.yak con. GE24787*	*D.ere con. GG24235*	*D.ana con. GF15002*	*D.pse con. GA18970*	*D.pse con. GA25629*	*D.per con. GL14516*	D.wil con. GK14607	D.wil con. GK18077	D.mo jcon. GI17338	D.vir con. GJ16066	D.gri con. GH25261
c2e6oA	% Confidence	98.6	98.6	98.1	98.5	98.5	99.1	97.9	97.8	97.5	98.1	98.6	98.2	98.7	98	98
	% identity	15	12	13	13	16	16	20	11	14	11	15	13	16	17	19
Info:	% Coverage	96	96	92	96	96	89	94	94	92	94	96	88	96	92	92
a b c d	Residues	2-56	2-56	2-54	2-56	2-56	6-56	1-54	1-54	1-51	1-54	2-56	2-57	2-56	2-54	2-54
c2cs1A	% Confidence	98.6	98.5	98.1	98.5	98.4	99.1	97.9	97.7	97.6	98.1	98.6	98.3	98.6	98	98
	% identity	24	21	17	18	27	22	24	13	18	13	22	19	16	15	19
Info:	% Coverage	96	91	92	96	78	96	78	94	79	94	96	83	96	92	92
e f g	Residues	2-56	2-59	2-54	2-56	10-54	2-56	10-54	1-54	6-51	1-54	2-56	2-54	2-56	2-54	2-54
d1v64a	% Confidence	98.6	98.6	98.3	98.5	98.5	99.1	98	97.8	97.5	98.1	98.6	98.3	98.6	98.1	98.1
	% identity	17	16	15	15	21	17	19	19	17	19	15	19	17	17	19
Info:	% Coverage	91	79	91	91	91	92	91	91	88	91	91	90	91	91	91
j	Residues	3-54	10-59	3-54	3-54	3-54	3-55	3-54	3-54	2-50	3-54	3-54	3-59	3-54	3-54	3-54

a -transcription
b - cell cycle
c - hmg box-containing protein 1
d- solution structure of the hmg box domain from human hmg-box2 transcription factor 1

e - dna binding protein
f - pms1 protein homolog 1
g- solution structure of the hmg domain of human dna mismatch2 repair protein
h - HMG - box

Table 4B. Detailed Analysis of Functional Groups found in Mst35Ba matches

Sample Matches Mst35Ba		*Mst35 Ba*	*Mst35 Bb*	*D.sim GD21981*	*D.sec GM14632*	*D.yak GE24787*	*D.ere GG24235*	*D.ana GF15002*	*D.pse GA18970*	*D.pse GA25629*	*D.per GL14516*	D.wil GK14607	D.wil GK18077	D.moj GI17338	D.vir GJ16066	D.gri GH25261
c2e6oA	% Confidence	99.5	99.5	99.5	99.4	97.9	99.5	99.5	-	99.4	99.5	99.1	60.7	96.9	98	99.2
	% identity	11	11	11	12	11	10	19	-	15	11	10	23	17	11	14
Info: a b c d	% Coverage	47	47	48	50	31	38	24	-	30	89	33	23	16	33	80
	Residues	69-138	69-138	69-140	66-140	135-204	117-194	67-149	-	2-63	1-76	139-215	23-62	215-262	129-200	12-87
d1v64a	% Confidence	99.5	99.5	99.4	99.3	98	99.5	99.3	22.5	99.5	99.3	99	69.4	97.2	98.1	99.1
	% identity	13	13	13	12	17	11	15	20	11	18	12	8	18	14	13
Info: e	% Coverage	47	46	48	50	23	38	19	12	34	84	33	23	17	33	81
	Residues	169-138	71-138	73-144	66-140	153-204	117-195	83-149	1-71	3-73	5-76	139-215	23-62	215-264	129-200	10-86
c1hmfA	% Confidence	99.4	99.4	99.4	99.2	97.6	99.4	99.3	6.4	99.4	99.3	98.8	61	96.6	97.6	99
	% identity	18	20	16	13	21	18	21	32	13	21	27	23	24	30	26
Info: f g h	% Coverage	44	44	46	46	23	32	21	5	32	71	34	23	17	21	69
	Residues	74-139	74-138	76-144	76-144	153-204	129-194	78-149	32-65	3-69	15-75	138-215	23-62	215-264	149-195	21-86

a-transcription
b-cell cycle
c- hmg box-containing protein 1
d- solution structure of the hmg box domain from human hmg-box2 transcription factor 1

e - HMG-box
f-DNA-binding
g-high mobility group protein fragment-b
h-structure of the hmg box motif in the b-domain of hmg1

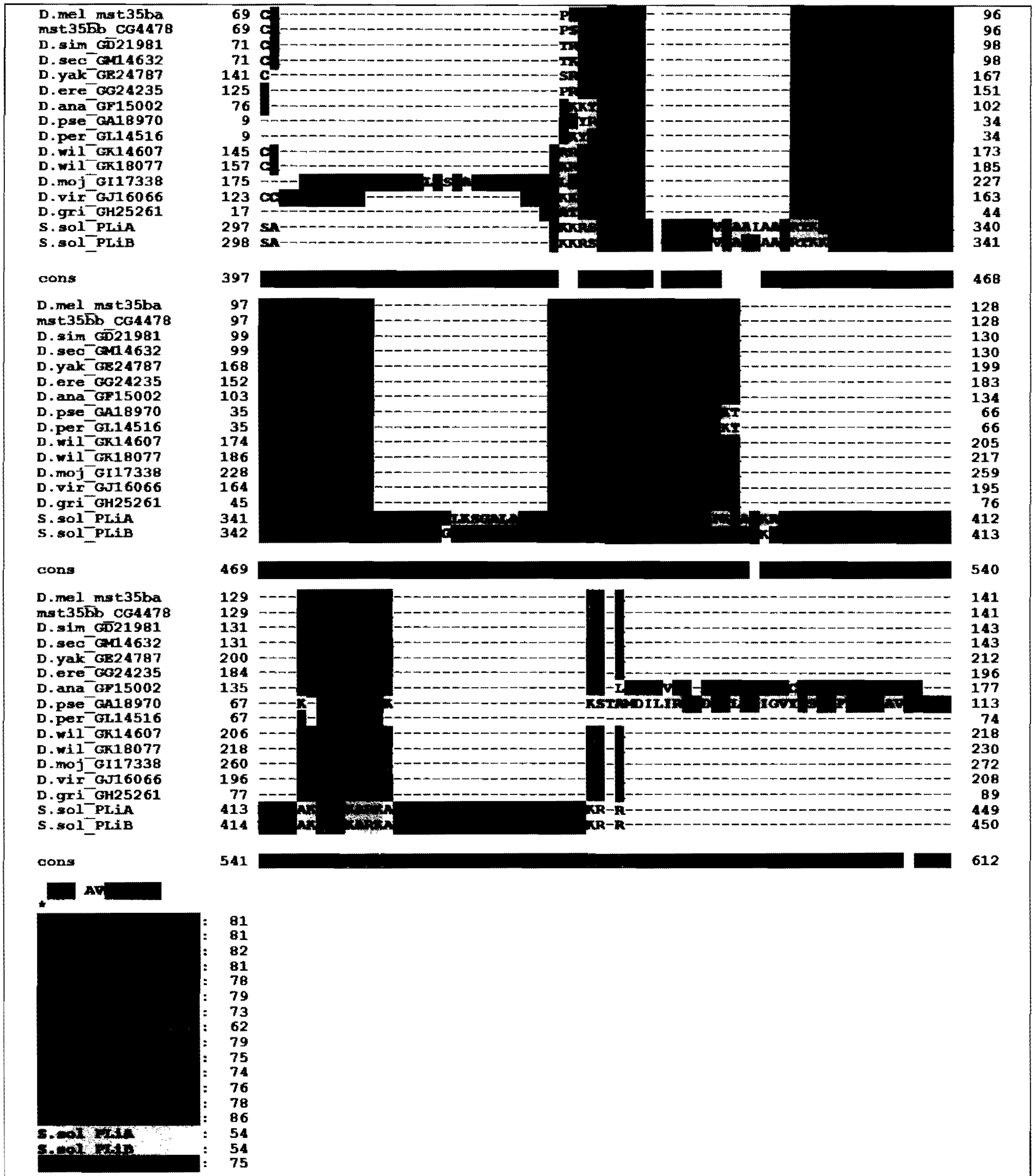
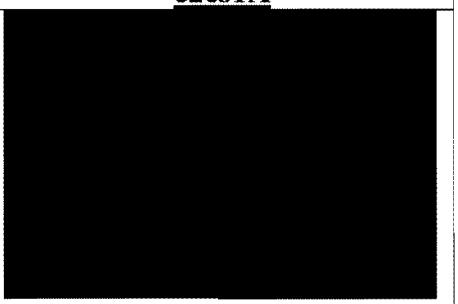
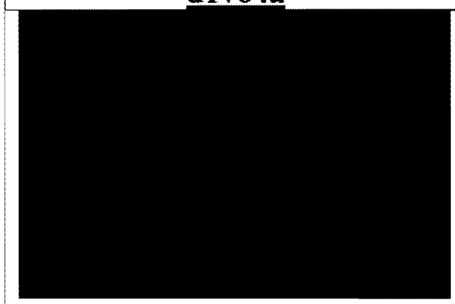
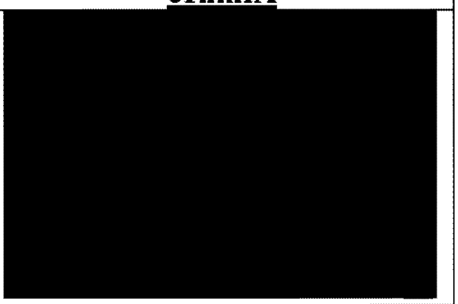


Figure 17. T-Coffee alignment of Drosophila protamine-like proteins and two protamines like proteins from Arctic surf Clams. The conserved sequence is found in the surf clam as well as in the protamine-like protein matches for the 12 sequenced Drosophila flies.

Figure 18: Secondary structures in wire frame for Mst35Ba functional groups

c2e6oA		**c2cs1A**
	c2e6oA - Involved in transcription and cell cycle. Also it is part Human HMG Box transcription factor 1. cs2cs1A - involved in DNA binding and is part of HMG.	
d1v64a		**c1hmfA**
	d1v64a - part of HMG Box. c1hmfA - involved in DNA binding. Also it is part of HMG protein.	
Refer to Tables 4A and 4B for more details for each functional group		

IV. PCR and Sequencing Analysis

Qiagen Kit (QIAMp) was used to isolate genomic DNA from two *D. pseudoobscura* flies following manufacturer's protocol. This extracted DNA was analyzed on 1% agarose gel as indicated in Figure 19 with its number of nucleotide bases being larger than 20,000 base pairs.

Using the designed primers through NCBI Primer BLAST and IDT Primer Quest^{SM} based on the transcript (mRNA) sequences as indicated in Table 5, we tried to isolate DNA from *D. pseudoobscura* and analyze it on a 2% agarose gel as indicated in Figure 20. After sequencing the PCR products that appeared to work, it was acknowledged that only two of the six primers were partially successful in extracting the region of interest of *D. pseudoobscura* GA18970. The primers that partially worked were Dpse35baMSP001F: CTTCCACGGCCGCCATCCAG, Dpse35baMSP001R: GCCTCCAGCAGTCGCACGTT and Dpse35baEWRP002F: TGCAGCTGTGGACGCCCTTG, Dpse35baEWRP002R: TGCGCGGTGGCATTTTTGGC.

Primer Dpse35baMSP001 sequence was only acquired through nucleotide BLAST two, which allows the aligning of two sequences. The expected primer locations and sequenced regions that were observed for each working primer location is illustrated in Figure 21. The following sequence was able to be extracted from *D. pseudoobscura* GA18970 through nucleotide BLAST for Dpse35baMSP001 with the sequence in pink and green indicating the scores:

```
CTTCCACGGCCGCCATCCAGCAACAAGTGTCCAACACCACGGAGGAGTCGG
CCCCGCCCACTCCCACCTCGCCCACTGGCAGCCTCAAGTCGAACACCCTGTC
CCTGTGCTCCACCGCTTCCGTGCCCATCGCCCGATCGGACAAGCACGTCGTC
CTGGCCATGCACCCCACGCGTCCCCGCGTCTCCTCCATGAACGCCAAGTTGG
ATCACACCAAAATCGACATGACCCTCTACAGAAGCCACGCTCAGCCAAAGA
CCCTGGACCCCGCTCCGGCCATCGAAGTGCGGGGAAATCTGCACGTGGGCA
TCAGCTACGATCCTGTGGGGGGTCTGCTCAACGTGCGACTGCTGGAGGC
```

A total of 270 base pairs of the total 358 base pairs were extracted. (Fig. 21)

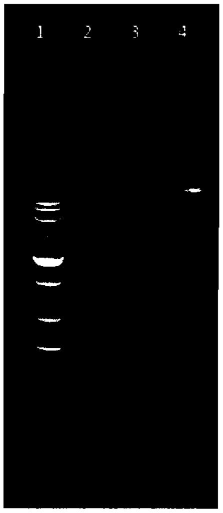
The following sequence was able to be extracted from *D. pseudoobscura* GA18970 (Fig. 21) through Mega BLAST for Dpse35baEWRP002 with the sequence in the red indicate a score above 200:

```
TGCAGCTGTGGACGCCCTTGAGCTCTGCCTCGGCCCAGGACATGAAAGTGGA
TTTGGGGGACATAATGGTGTCCCTGGCCTACCTGCCCTCGGCCGAACGCCTG
ATGGTGGTGCTGATCAAGGCCAGAAATCTGCGGATTGTGGACGATGCCAGG
AACTCCTCCGATCCGTACGTGAAGGTGACTCTCCTCGGGCCTGTGGGCAAGA
AAATGAAGAAGCGCAAGACCGGCGTCCAGCGGAGCACCGTCAATCCTGTGT
ACAACGAGGCCCTGGCCTTTGATGTCAACAAGGAGACGCTGAAGAACTGCG
TGCTCGAGTTTACTGTCGTCCACGACGGTCTTTTGGGATCGAGCGAAATATT
GGGCCGCACTCTCATCGGCAACTCGTCCGAGGTGCGCACTGAGGAGAAGAT
CTTCTTCGAGGAGATGTTTCGCGCCAAAAATGCCACCGCGCA
```

The extracted genomic size for Dpse35baEWRP002 was 302 base pairs, which is comparable and close to the expected base pair size of 454 base pairs.

Table 5. Primers for *D. pseudoobscura* gene transcript location of GA18970 (GI: 198475489)

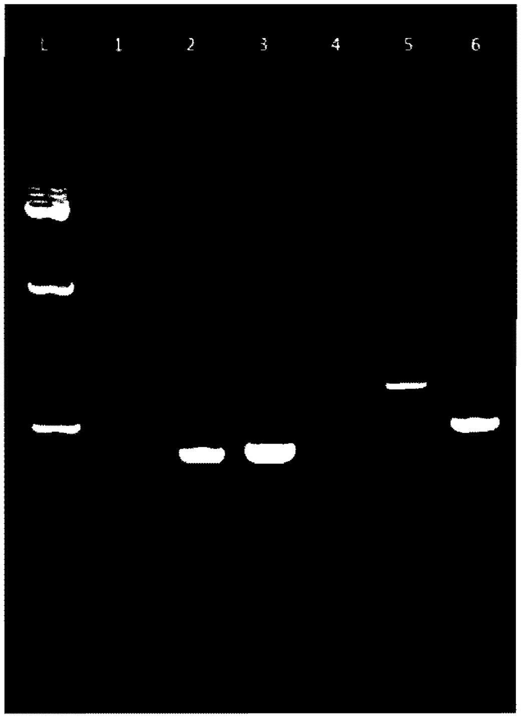
Prime Name	Primer Sequence	Primer Set #	Size (bp)
Dpse35baSshRP002F	CCTGTCGCCCCGGGAGATGA	2F	434
Dpse35baSshRP002R	GCACCCAGAAGGCCGGATCG	2R	
Dpse35baSrlRP002F	GCAGAAGGCGAGCTTCCGCA	4F	505
Dpse35baSrlRP002R	AGCGGTGGAGCACAGGGACA	4R	
Dpse35baMSP001F	CTTCCACGGCCGCCATCCAG	5F	358
Dpse35baMSP001R	GCCTCCAGCAGTCGCACGTT	5R	
Dpse35baMMESP002F	CTTCCACGGCCGCCATCCAG	6F	582
Dpse35baMMESP002R	TGCACACGTGCCGCGAGTAG	6R	
Dpse35baEWRP001F	TGCAGCTGTGGACGCCCTTG	8F	460
Dpse35baEWRP001R	ACCCATTGCGCGGTGGCATT	8R	
Dpse35baEWRP002F	TGCAGCTGTGGACGCCCTTG	9F	454
Dpse35baEWRP002R	TGCGCGGTGGCATTTTTGGC	9R	



Lane	Sample
1	1 Kb ladder
2	Negative control
3	Pse#1 DNA
4	Pse#2 DNA

Pse#1 and Pse#2
contain 2 Dpse flies each
1% Agarose Gel

Figure 19. DNA Extraction of *D. pseudoobscura*



Lane	Primers
L	1 Kb ladder
1	Dpse35baSshRP002 #2
2	Dpse35baSrlRP002 #4
3	Dpse35baMSP001 #5
4	Dpse35baMMESP002 #6
5	Dpse35baEWRP001 #8
6	Dpse35baEWR002 #9

Pse#1 was used for PCR,
2% Agarose Gel

Figure 20. PCR Gel Results for *D. pseudoobscura GA18970*

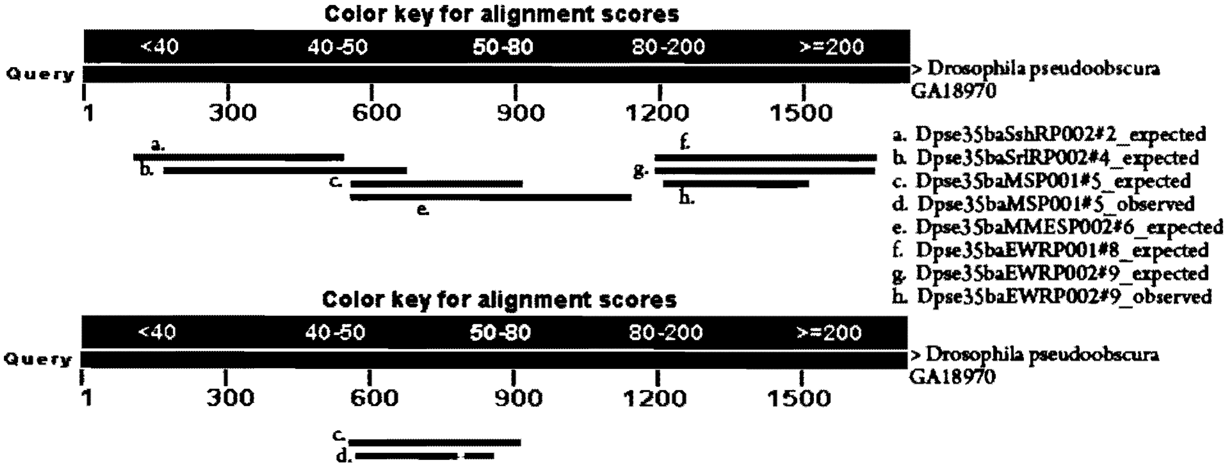

Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
> Drosophila pseudoobscura GA18970
1
300
600
900
1200
1500
a.
b.
c.
e.
f.
g.
h.
a. Dpse35baSshRP002#2_expected
b. Dpse35baSrlRP002#4_expected
c. Dpse35baMSP001#5_expected
d. Dpse35baMSP001#5_observed
e. Dpse35baMMESP002#6_expected
f. Dpse35baEWRP001#8_expected
g. Dpse35baEWRP002#9_expected
h. Dpse35baEWRP002#9_observed
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
> Drosophila pseudoobscura GA18970
1
300
600
900
1200
1500
c.
d.

Figure 21. Sequenced regions of *D. pseudoobscura* GA18970 versus expected regions

V. Transmission electron microscopy

Using TEM samples made from virgin adult *D. pseudoobscura* flies, we analyzed and identified chromatin condensation in the sperm nucleus during nuclear transformation. Figure 22 shows several stages of a nuclear transformation in sperm nuclei.

We visualized the progression of chromatin condensation within the sperm nuclei of *D. pseudoobscura* by transmission electron microscopy (TEM). Figure 22 shows the patterns of chromatin condensation in elongated spermatids during nuclear transformation. The chromatin appears diffuse in the early stages of nuclear transformation (Figs. 22A and 22B). As the nucleus becomes more condensed in later stages of transformation, chromatin patterning becomes evident (Figs. 22C – 22F). As the chromatin approaches full condensation, regions that appear to be voids are visible within the nuclei (Figs. 22G and 22H). The sperm shown in Figures 22A – 22H were from the basal end of the *D. pseudoobscura* testis. Figures 22I and 22J show mature sperm nuclei within the seminal vesicle. The chromatin is assumed to be fully compacted at this stage

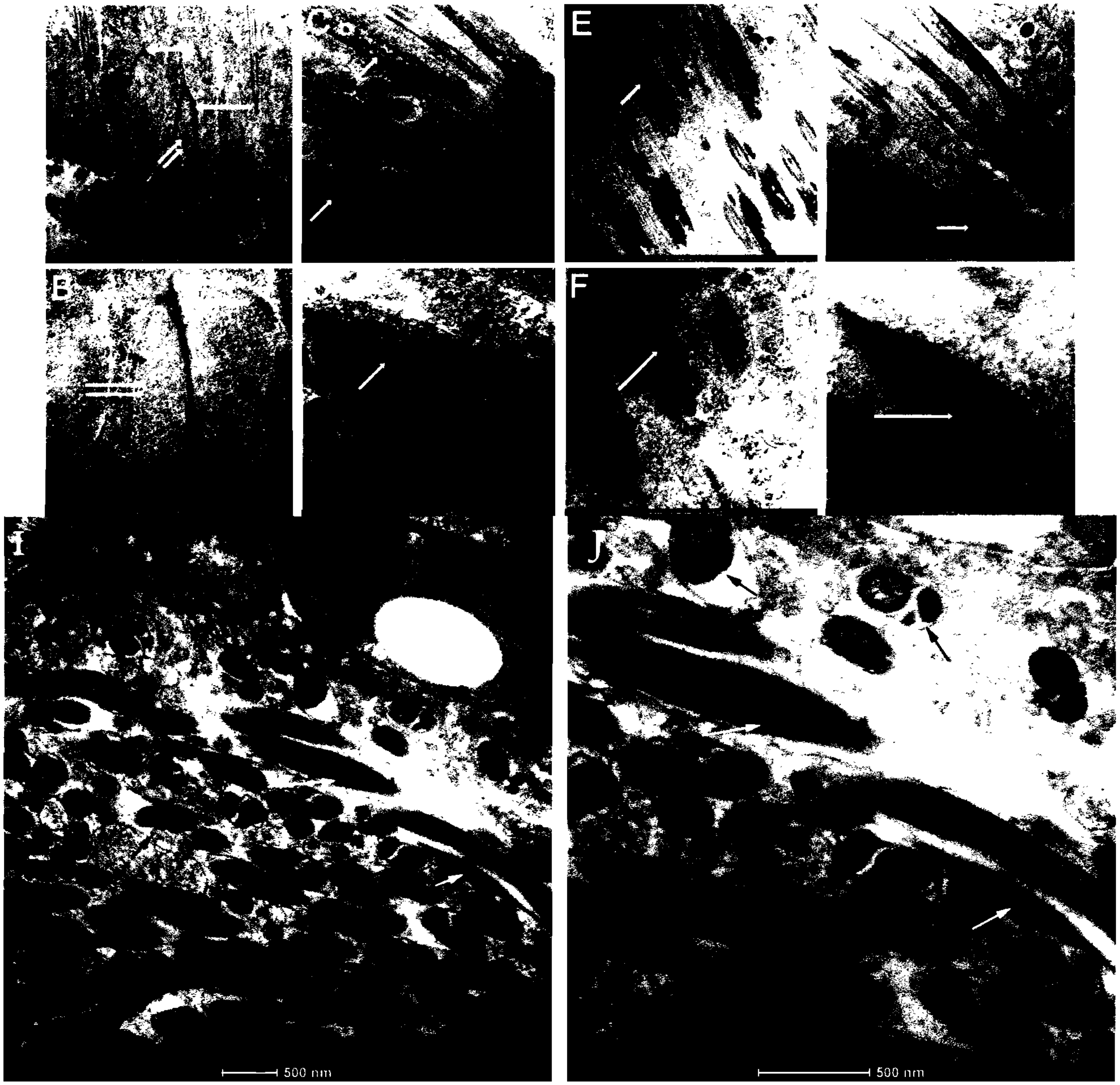


Figure 22. Transmission electron micrographs of condensing sperm chromatin in *D. pseudoobscura*. (A, C, E, G, I) low magnification views of chromatin becoming increasingly more condensed, left to right. (B, D, F, H, J) Higher magnification views of the sperm nuclei. In panels A and B, single arrows = developing acrosome and double arrows = chromatin. In panels C – J single arrows = condensing chromatin.

Discussion

I. Putative translational expression regions among the 12 sequenced Drosophila species

Our results indicate that the best nucleotide transcript and genomic DNA sequence matches for Mst35Ba and Mst35Bb among the 12 Drosophila species are identical with the exception of *D. grimshawi* (Figs. 4A and 4B). This similarity, as stated earlier, is due to a duplication event of Mst35Ba (Raja et al. 2005; Birkhead et al. 2009). The 3' untranslated region (UTR) region that is found in mice is known to be involved in translational repression after being mutated (Raja et al. 2005; Zhong et al. 2001). In contrast, the promoter region and the 5' UTR for Mst35Ba and Mst35Bb have a high identity between each other and are responsible for translational repression after gradual 5' upstream deletions to create a mutant line (Raja et al. 2005). The high identity match between Mst35Ba and Mst35Bb at the 5' end can be attributed to the putative conserved region among the 12 Drosophila species (Figs. 10A and 10B). NCBI ORF finder indicated that the conserved region for each Drosophila fly was within the open reading frame that occurs near the 5' end of each respective genomic DNA sequence. *D. grimshawi* GH12778 was the only genomic DNA match that had a similar conserved region at the 5' end of its sequence. However, the overall length of this region was increased to 113 amino acids as compared to approximately 56 and 62 amino acids respectively for the conserved region of Mst35Ba and Mst35Bb. The other genomic DNA match, *D. grimshawi* GH13870, did not appear to have a conserved region when compared to the rest of the matches. Moreover, the overall consensus among the conserved regions in the 12 Drosophila species was 95% with *D. grimshawi* GH12778. This high consensus score correlates with that the 5' region will have high identity score. Overall, the putative conserved regions and their respective 5' open reading frames are predicted to be involved in transcriptional expression. Alternatively, if the conserved regions and their respective 5' open reading frames could be involved in transcriptional repression if modified.

II. Presence of conserved region in protamine-like proteins in other organisms.

The best-matched protamine-like proteins among the 12 Drosophila species for Mst35Ba and Mst35Bb were compared with other protamine-like proteins from other unrelated species for a similar conserved region. Interestingly, *Spidsula solidissima,* an arctic surf clam has similar conserved region such as the putative conserved region that was predicted in Figures 10A and 10B. In contrast, *Mullus surmuletus*, which only has one protamine-like protein, does not have a highly conserved domain as *S. solidissima* and the all the Drosophila matches. The interesting aspect of these preliminary findings is that the conserved domain seems to only be found in species that have two protamine-like proteins.

III. Phylogenetic relationship of the 12 sequenced Drosophila species

Our results indicate that the matches for the species that are in the melanogaster sub group (*D. melanogaster, D. simulans, D. erecta, D. yakuba*) shows consensus to the established phylogenetic tree (Fig. 1) by being conserved and identical in the phylogenetic trees for the protein and genomic DNA matches. Interestingly the branching patterns in the phylogenies generated using Mst35Ba and Mst35Bb match with the relationships that have been established in the phylogenetic tree in Figure 1. In contrast, the nucleotide transcript matches illustrate only *D. simulans, D. sechelia,* and *D. melanogaster* have the same phylogenetic relationship as the established phylogenetic tree (Fig. 1). Also the phylogenetic relationship for *D. grimshawi* indicates that it evolved from the Drosophilidae Family as a separate lone group as shown in Figures 7A and 7B instead of branching from repleta and virilis sister groups.

As Mst35Bb emerged due to a duplication event of Mst35Ba (Birkhead et al. 2009; Raja et al. 2005), this could be a reason as to why the majority of the distant species from Mst35Bb matches have a larger query coverage with lower E-values when compared to the matches that have been found for Mst35Ba for the nucleotide transcript and genomic DNA matches (Figs. 4A, 4B, 5A, and 5B). Therefore the generated phylogenetic trees for nucleotide transcript and genomic DNA matches for Mst35Bb indicate a greater variance when compared to Mst35Ba matches.

IV. Amino acid analysis

There have been numerous studies conducted on the number of amino acids present and their respective percentages for histone H1 linker like proteins, protamine-like proteins, and true protamines. (Eirin-Lopez et al. 2009; Eirin-Lopez et al. 2006b; Birkhead et al. 2009; Balhorn et al. 2007) Protamine-like proteins evolved into protamines due to a separation of small arginine-rich regions that occurred early in the evolution of these proteins (Eirin-Lopez et al. 2009). The protamine-like proteins evolved from an H1 histone lineage (somatic H1, RD, H1, R1, and SNBPs) and belong to the same monophyletic group (Eirin-Lopez et al. 2009; Eirin-Lopez et al. 2006b), which explains the similar number of different amino acid percentages between histone H1 linker proteins and protamine-like proteins. A study with protamine-like proteins PLiA and PLiB from *S. solidissima* was compared to a true protamine found in *D. labrax* in terms of the percentage and the total number of amino acids in comparison (Saperas et al. 1993). Interestingly, this study found that the concentration of serine, lysine, and arginine in *S. solidissima* was similar to the matches found for Mst35Ba and Mst35Bb (Saperas et al. 1993). In general, the combination or the ratios of lysine and arginine amino acids that are in protamine-like proteins are important indicators for binding DNA. In addition, the high percentage of alanine and serine amino acids is a characteristic of protamine-like proteins (Saperas et al. 1993).

In the whole protein matches amino acid percentage breakdown, all matches have approximately 3 to 9 % cysteine for the conserved regions. The importance of cysteine is that it is able to form disulfide bonds to increase sperm chromatin compactness (Cheng et al. 2009; McBride et al. 1992). Both whole proteins and the conserved protein regions had a high percentage of lysine and arginine amino acids. The common importance of arginine and lysine is that they are basic amino acids that have positive charge at physiological pH. The higher percentage of these amino acids means that the protamines-like proteins use them to increase their affinity with the DNA during chromatin condensation. Furthermore, the arginine has the higher hydrogen bonding potential than lysine, which protects the condensing chromatin from DNA damaging agents.

V. Putative Conserved as DNA Binding regions - DNA Binder, BindN+, BindN-RF

DNA-Binder (http://www.imtech.res.in/raghava/dnabinder/) was used to predict that the conserved regions may be a DNA-binding domain based on support vector machine (SVM) models. DNA-Binder uses this for classification of the protein based on a regression algorithm models to predict inputted amino acid sequence based on a user-defined threshold is a DNA binding protein or a non-DNA binding protein. The three-dataset types are realistic, alternative, and main set. As the name states, the realistic dataset compares the amino acid sequence as it would be in nature with the 1:10 (DNA-binding to non-DNA-binding protein chains). Realistic Dataset[1] has the following parameters: sensitivity set to 47.95%, specificity set to 93.33% and accuracy set to 89.31%. Additionally the realistic dataset searches through 146 DNA-binding protein chains and 1500 non DNA-binding chains.

The alternative set compares the whole library of DNA-binding and non-DNA-binding protein chains. Alternative Dataset[2] has the following parameters: sensitivity set to 72.51%, specificity set to 72.33%, and accuracy set to 72.42%. Additionally the alternative dataset searches through a wider range of DNA-binding and non DNA-binding protein chains as compared to the realistic dataset. This range includes 1153 DNA-binding proteins and 1153 non DNA-binding protein chains. The alternative dataset is usually used to analyze full-length protein sequences. The combination of alternative dataset and realistic dataset was used to analyze protein BLAST matches for SNBP.

Lastly, the main set is used specifically to identify domains within large protein sequences to be DNA binding or non-DNA binding. Main Dataset[3] has the following parameters: sensitivity is set to 78.11%, specificity is set to 80.80%, and accuracy is set to 79.80%. The main data set searches through 146 DNA-binding protein and 250 non-DNA binding chains. The purpose of the main dataset is to identify and search domain sequences within larger protein sequences for their likeliness to be DNA-binding regions. Hence, this was used to analyze the putative DNA binding domains matches for each SNBP. If the score is greater or close to one then the likely chance of it to be DNA binding domain are high. In contrast, if the score is closer to -1 or less than the amino acid sequence is more likely to be non-DNA binding domain. If the number is near zero and in between -1 and 1 then it could be DNA binding domain or non-DNA binding domain (Kumar et al 2007). In general, the majority of the matches were DNA-Binding proteins with some minor divergences due to their low query convergence and increased variance in the amino acid distribution due to their length (Table 2).

In BindN+ the amino acid sequence is analyzed and predict based upon two Protein Data Bank (PDB) datasets (PDNA-62 and PRINR25) (Wang et al 2010). The BindN-RF uses a

Random Forest algorithm to predict the DNA binding residues. The user-defined amino acid sequence is searched through PDB PDNA-62 database. Overall, BindN-RF is able to achieve higher accuracy compared to BindN+ (Wang et. al 2009). Additionally, BindN and BindN+ search for the evolutionary information of the amino acid sequence by having the amino acid sequence be searched three times against the UniPortKB database.

In Figures 13A, 13 B, 14A, and 14B the BindN+ (relaxed) and Bind-RF (strict) indicate conserved region (darkly shaded) contains several DNA binding residues, which are analogous to conserved matches in the 12 sequenced Drosophila flies for each respective SNBP. The darkly shaded region in black is the conserved region in both Mst35Ba (Prot A) and Mst35Ba (Prot B). The darkly shaded region in blue is the conserved region in only Mst35Bb (Prot B). The conserved regions in the SNBP matches all contain a high concentration of DNA binding residues. The Bind N+ specificity was set 79% as recommended. Likewise, the specificity for BindN-RF was set to the recommended value of 78.22% (http://bioinfo.ggc.org/; Wang et al. 2009).

VI. Functional Groups

After determining that the different matched sequences for each species conserved region is a putative DNA-binding domain, we searched for the function of these conserved regions and the whole proteins using Swiss Model Interpro Domain scan and Phyre2. The Swiss Model Interpro Domain scan was able to search and identify the regions of a sequence that belong to particular protein domains, superfamilies, and families. The Swiss Model Interpro Domain uses HMMPFam, which is a collection multiple sequence alignment of Hidden Markov models that cover many commonly known protein domains and families; HMMTgr is a collection of protein families that have been organized and collected by multiple aligned sequences that identify the functionality of related proteins based on the homology of the sequence; ProfileScan is able to identify significant sites, patterns of known protein families; Superfamily is a library of hidden Markov models that are representative of proteins of known functions; ProDom is a large collection of homologous domains where recursive PSI-BLAST is conducted to analyze the domain arrangement between the protein sequence and their families. FPrintScan searches the conserved motifs that help characterize the protein family; HMMSmart is able identify and annotate the genetic mobile domains and analyze their domain architecture; and ScanRegExp is a database of protein families and domains that are composed of biologically important sites and patterns (Zdobnov et al. 2001).

The homologous regions are gathered together and converted to Hidden Markov Models (HMM). The HMM is able to capture the mutations that have occurred through the evolutionary time of the sequence. Therefore the HHM is able to act as an evolutionary fingerprint for the protein's evolutionary history. Additionally, the 3D protein structures for the protein are generated by extracting the protein sequence of the known approximately 65,000 3D protein structures and then running PSI BLAST to generate HMM for a sequence of a known structure. This is then made into HMM Database of Known Structures. The user defined protein sequence is scanned and matched through the HMM Model Database of Known Structures, which yield an alignment that can be interpreted through high confidence score, coverage, and identity match. Then finally the alignment is used to create a 3D model of the user-defined protein sequence (Kelley et al 2009).

The functional groups present in Table 4B are present in all conserved regions. Additionally, these functional groups are present in nearly all Drosophila fly protein matches with the exception of *D. pseudoobscura* GA18970, which can probably be attributed to its large number of amino acids. The functional groups listed in Tables 4A and 4B belong to the high mobility group (HMG) box, which has been reported to be a DNA-binding domain and is involved in transcription (Qin et al. 2003). As expected the, Swiss-Model InterPro Scan found large coverage of the HMG box with partial coverage of the DUF1074 family of proteins, whose function is unknown. DUF1074 is part of HMG box like superfamily that contains six family members (CHDNT, DUF1014, DUF1074, DUF1898, HMG box, and YABBY) as annotated by the Sanger Institute (Bateman et al. 2004). The functional group of 3fghA (not shown) is a known DNA binding subunit that has excellent confidence above 98% (Pearl et al. 2005). All of the conserved regions (Figs. 10A and 10B) contain an expanded overlap of the DUF1074 protein family of unknown function and HMG. Only *D. pseudoobscura* matches and the controls for Mst35Ba and Mst35Bb have been shown in Figure 15. Thus, there is an strong possibility that the HMG group and DUF1074 could be involved in DNA-binding and the chromatin condensating process.

Lastly, the conserved region (Figs. 10A and 10B) appears to be almost identical to each other among the 12 sequenced Drosophila flies in terms of their secondary wire frame structure. In addition, the consensus secondary wireframe structures appear to have similar shape to known secondary wireframe structures in terms of the three helices of known HMG boxes.

VII. Spermiogenesis, Chromatin Condensation and Nuclear Transformation

The condensation of sperm chromatin is a process that occurs during spermiogenesis and nuclear transformation. Our preliminary work on *D. pseudoobscura* transforming nuclei shows that we can visualize successive stages of chromatin condensation by TEM. During spermiogenesis, the histones are replaced by protamines (Kasinsky et al. 2011). Throughout spermiogenesis, the chromatin is able to condense and become a stable and compact structure due to this exchange (Kasinsky et al. 2011). In contrast to protamine: DNA interactions, histones compact DNA wrapping the DNA molecule two and a half times. H1 linker histones bind to DNA that connects adjoining nucleosomes (Kasinsky et al. 2011; van Holdie et al. 1998). Protamines bind directly to the major groove of DNA by interacting with the phosphates of the backbone (Eirin-Lopez et al. 2009; Kasinsky et al. 2011). In the events just prior to protamine displacement, histones become acetylated in vertebrates and invertebrates, which lowers the histone and DNA interaction and increases the protamine displacement of the histones (Kasinsky et al. 2011; Oliva et al. 1991).

During nuclear transformation when round spermatids undergo the transition into mature spermatozoa, sperm nuclear basic proteins replace histones and the majority of nucleosome structure is lost (Ward and Coffey et al. 1991; Ward et al. 2011). However, in some species, such as humans, some fraction of histone-bound DNA and nucleosome structure is retained (van der Heijden et al., 2006; van der Heijden et al., 2008; Vavouri and Lehner, 2011). Chromatin condensed with P type, PL type, and H type proteins give rise to a variety of chromatin patterns including lamellar and fibrogranular (Caceres et al., 1999; Harrison et al., 2005; Kasinsky et al., 2011; Eirin-Lopez et al., 2011; Saperas et al., 1993).

Mammalian sperm nuclear shape is disrupted if protamine expression is abnormal, with the heads assuming a enlarge, rounded shape instead of a paddle-like flattened shape in humans (Balhorn et al., 1988). Problems with protamine expression is often associated with male infertility (Oliva, 2006). Other authors have suggested that nuclear shaping and sperm head shape is associated with SNBPs (Ausio, et al., 2006; Martin-Coello et al., 2009). Ultimately, we will test the hypothesis that the variable protamine-like proteins identified in the current work are involved in variable chromatin patterning in the 12 sequenced Drosophila species.

Similarly, we predict that variable chromatin patterning is involved in achieving the species-specific shape of the sperm nuclei.

Future Studies

Our work strongly suggests that the homologues for Mst35Ba and Mst35Bb are present in all of the currently sequenced Drosophila species. Additionally, there appears to be a conserved DNA binding domain present in these proteins. The next step in this work will be to continue PCR analysis of the putative sequences found in the current study, as well as continue the analysis of chromatin condensation patterns for the 12 Drosophila species. Furthermore, analyze the chromatin condensation patterns of the 12 Drosophila species with the relationship to the concentration of arginine and lysine present for the respective matches. Future studies may also use the data from the current work as a starting point to generate mutant flies among 12 sequenced Drosophila flies. The development of mutant flies will aid in the better understanding of how these proteins affect nuclear transformation and chromatin condensation during spermiogenesis. As our lab has already developed an *in vitro* cyst culture for *D. pseudoobscura,* we hope to use these new mutant flies to study spermatogenesis. Additionally, we will perform a 12-species analysis of the other SNBP (Mst77F) found in the *D. melanogaster* sperm nucleus. Overall, a better understanding of fertility and the role of these particular protamines in the development of mature sperm will be achieved.

Literature Cited

Arnold, K., Jürgen K., Torsten S., and Lorenza B. "The SWISS-MODEL Workspace: A Web-based Environment for Protein Structure Homology Modeling." *Bioinformatics,* 22, (2006): 195 - 201.

Ausio, J. "Histone H1 and Evolution of Sperm Nuclear Basic Proteins." *Journal of Biological Chemistry*, 274, (1999): 31115 – 31118.

Balhorn, R."The protamine family of sperm nuclear proteins." *Genome Biology* 8.9 (2007): 1 - 8.

Balhorn, R., Reed, S., and Tanphaichitr, N.. "Aberrant protamine 1/protamine 2 ratios in sperm of infertile human males." *Experientia,* 44, (1988): 52 - 55.

Bateman, A., Coin, L., Durbin, R., Finn, RD., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, EL., Studholme, DJ., Yeats, C., and Eddy, SR. "The Pfam protein families database." *Nucleic Acids Research,* 32, (2004): D138 – D141.

Birkhead, T. R., Hosken, D. J., and Pitnick., "S. *Sperm Biology - an Evolutionary Perspective*." 1st ed. San Diego, Ca: Academic Press, (2009).

Caceres, C. and et al. "DNA-interacting proteins in the spermiogenesis of the mollusc Murex bandaris." *Journal of Biology Chemstry,* 274, (1999): 649 - 56.

Cheng, W., An, L., Wu, Z., Zhu, Y., Liu, J. Gao, H., Li, X., Zheng, S., and Tian, J. "Effects of Disulfide Bond Reducing Agents on Sperm Chromatin Structural Integrity and Developmental Competence of in Vitro Matured Oocytes after Intracytoplasmic Sperm Injection in Pigs." *Reproduction,* (2009): 633 - 643.

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

Dorus, S., Freeman, ZN., Parker, ER., Heath, BD., and Karr, TL. "Recent Origins of Sperm Genes in Drosophila." *Molecular Biology and Evolution,* 25.10, (2008): 2157 - 2166.

Gilbert, LI. "Drosophila Is an Inclusive Model for Human Diseases, Growth and Development." *Molecular and Cellular Endocrinology,* 293, (2008): 25 - 31.

Eirin-Lopez, J. M., and Ausio, J. "Origin and evolution of chromosomal sperm proteins." *BioAssays,* 21, (2009): 1062 - 1070.

Eirin-Lopez, J. M., Frehlick, L. J., and Ausio J. "Protamines, in the Footsteps of Linker Histone Evolution." *The Journal of Biological Chemistry,* 281, (2006): 1 - 4.

Eirin-Lopez, J. M., Lewis, J. D., Howe le, A. and Ausio, J. "Common phylogenetic origin of protamine-like (PL) proteins and histone H1: Evidence from bivalve PL genes." Molecular *Biology and Evolution*, 23, (2006b): 1304 – 1317.

Fuller, M. T. "Genetic Control of Cell Proliferation and Differentiation in Drosophila Spermatogenesis." *Cell and Developmental Biology,* 9, (1998): 433 - 444.

Hammoud, S. S., Nix, D. A., Zhang H., Purwar J., Carrell, D. T., and Cairns, B. R., "Distinctive Chromatin in Human Sperm Packages Genes for Embryo Development." *Nature,* 460, (2009): 473 - 78.

Harrison, L. G., et al. "Possible mechanisms for early and intermediate stages of sperm chromatin condensation patterning involving phase separation dynamics." *Journal of Experimental Zoology,* 303A, (2005): 76 - 92.

Joly D., Bazin, C., Zeng, L-W., and Singh. R. "Genetic basis of sperm and testis length differences and epistatic effect on hybrid inviability and sperm motility between *Drosophila simulans* and *D. secheillia*." *Heredity,* 78, (1997): 354 - 362.

Kasinsky, H. E., Eirin-Lopez, J. M., and Ausio, J. "Protamines: Structural Complexity, Evolution and Chromatin Patterning." *Protein & Peptide Letters,* 18, (2011): 1 - 17.

Kelley, L.A., and Sternberg M.J.E. "Protein structure prediction on the web: a case study using the Phyre server." *Nature Protocols,* 4, (2009): 363 - 371.

Kumar, M., Gromiha, M. M., and Raghava, G. P. S. "Identification of DNA-binding proteins using support vector machines and evolutionary profiles." *BMC Bioinformatics,* 8:463, (2007): 1 – 10.

Lee, C., Grasso, C., and Sharlow. M.F. "Multiple Sequence Alignment Using Partial Order Graphs." *Bioinformatics,* 18, (2002): 452 - 64.

Lu, L.Y., Wu, J., Ye, L., Gavrilina, G.B., Saunders, T.L., et al. "RNF8-dependent histone modifications regulate nucleosome removal during spermatogenesis." *Developmental Cell,* 18, (2010): 371 - 384.

Markow, T. A., and O'Grady, P. M. "Drosophila Biology in the Genomic Age." *Genetics Society of America,* 177, (2007): 1269 - 1276.

Martens, G., Humphrey, E. C., Harrison, L.G., Silva-Moreno, B., Ausió, J., et al. "High-pressure freezing of spermiogenic nuclei supports a dynamic chromatin model for the histone-to-protamine transition." *Journal of Cellular Biochemistry,* 108, (2009): 1399 - 1409.

Martin-Coello, J., et al. "Sexual selection drives weak positive selection in protamine genes and high promoter divergence, enhancing sperm competitiveness." *Proceedings of the Royal Society Biological Sciences,* 276, (2009): 2427 - 2436.

McBride, A. A., Klausner, R. D., and Howley, P. M. "Conserved Cysteine Residue in the DNA-binding Domain of the Bovine Papillomavirus Type 1 E2 Protein Confers Redox Regulation of the DNA-binding Activity in Vitro." *Biochemistry,* 89, (1992): 7531 - 535.

Njogu, M., Ricketts, PG., Klaus AV. "Spermiogenic cyst and organ culture in Drosophila pseudoobscura." *Cell and Tissue Research,* 341 (2010): 453 - 464.

Noguchi, T., and Miller, K. G. "A role for actin dynamics in individualization during spermatogenesis in *Drosophila melanogaster. Development,* 130, (2003): 1805 - 1816.

Oliva, R. "Protamines and infertility." *Human Reproduction,* 12, (2006): 417 - 35.

Oliva, R., and Dixon, G.H. "Vertebrate protamine genes and the histone to-protamine replacement reaction." *Progress in Nucleic Acid Research and Molecular Biology,* 40, (1991): 25 - 94.

Pasini, M. E., Caviglia, O., Redi, A. C., and Perotti, E. M. "Ultrastructural and Cytochemical Analysis of Sperm Dimorphism in Drosophiloa Subobscura." *Tissue and Cell,* 28.2, (1996): 165 - 175.

Pearl, F., Todd, A., Silltoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J., and Orengo, C., Cuff, A., Ian, S., Tony, L., and Andrew C. "The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis." *Nucleic Acids Research, 33,* (2005): D247 - D251.

Qin, J., Kang, W., Leung, B., and McLeod, M. "Ste11p, a High-Mobility-Group Box DNA-Binding Protein, Undergoes Pheromone- and Nutrient-Regulated Nuclear-Cytoplasmic Shuttling." *Molecular and Cellular Biology,* 23.9, (2003): 3253 - 264.

Raja, S. J. "Chromatin Condensation during Drosophila Spermiogenesis and Decondensation after Fertilization." *Philipps-Universität Marburg -Entwicklungsbiologie-* (2005): Dissertation.

Raja, S. J., and Renkawitz-Pohl, R. "Replacement by Drosophila Melanogaster Protamines and Mst77F of Histones during Chromatin Condensation in Late Spermatids and Role of Sesame in the Removal of These Proteins from the Male Pronucleus." *Molecular and Cellular Biology,* 25.14, (2005): 6165 - 177.

Rathke, C. "Transition from a nucleosome-based to a protamine-based chromatin configuration during spermiogenesis in Drosophila." *Journal of Cell Science,* 120, (2007): 1689 - 700.

Richards, S., Yue, L., and Bettencourt, B. R. "Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution." *Genome Research, 15,* (2005): 1 - 18.

Ricketts, PG. A., Minimair M., Yates, R. W., Klaus, A. V., "The Effects of Glutathione, Insulin and oxidative stress on cultured spermatogenic cysts." *Spermatogensis*, 1 (2011): 159 - 171.

Rooney, A. P., Zhang, J., and Nei, M. "An unusual form of purifying selection in a sperm protein." *Molecular Biology Evolution,* 17, (2000): 278 - 283.

Saperas, N., Ribes, E., Garcia-Hegart F., and Chiva, M. "Differences in Chromatin Condensation during Spermiogenesis in Two Species of Fish with Distinct Protamines." *The Journal of Experimental Zoology,* 256.2, (1993): 185 - 94.

Tommaso, Paolo D., Sebastien Moretti, Ioannis Xenarios, Miquel Orobitg, Alberto Montanyola, Jia-Ming Chang, Jean-Francois Taly, and Cedric Notredame. "T-Coffee: a Web Server for the Multiple Sequence Alignment of Protein and RNA Sequences Using Structural Information and Homology Extension." *Nucleic Acids Research,* 39, (2011): W13 - W17.

Tweedie S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R.. Zhang, H., and The FlyBase Consortium. "FlyBase: enhancing Drosophila Gene Ontology annotations." *Nucleic Acids Research,* 37, (2009): D555 - D559.

van der Heijden, G. W., et al. "Sperm-derived histones contribute to zygotic chromatin in humans." *BMC Evolutionary Bioliology,* 8, (2008):.

van der Heijden, G. W., et al. "Transmission of modified nucleosomes from the mouse male germline to the zygote and subsequent remodeling of paternal chromatin." *Developmental Biology,* 298, (2006): 458 - 469.

van Holdie, K. E. "Chromatin." *Journal of Molecular Recognition,* 2, (1988):.

Vavouri, T. and Lehner, B. "Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome." *PLos Genetics,* 7, (2011).

Wang, L,, Yang, M. Q., Yang, J. Y,, and Huang, C. "BindN+ for Accurate Prediction of DNA and RNA-binding Residues from Protein Sequence Features." *BMC Systems Biology,* 4, (2010):.

Wang, L,, Yang, M. Q., and Yang, J. Y,. "Prediction of DNA-binding Residues from Protein Sequence Information Using Random Forests." *BMC Genomics,* 10, (2009).

Ward, W.S. "Function of sperm chromatin structural elements in fertilization and development." *Molecular Human Reproduction,* 16.1, (2011): 30 - 36.

Ward, W. S. and Coffey, D. S. "DNA packaging and organization in mammalian spermatozoa: comparison with somatic cells." *Biology of Reporduction,* 44, (1991) 569 - 574.

White-Cooper, H. "Studying how flies make sperm-investigating gene function in Drosophila testes." *Molecular and Cellular Endocrinology, 306,* (2009): 66 - 74.

Zdobnov, E.M., and Apweiler R. "InterProScan - an integration platform for the signature-recognition methods in InterPro." *Bioinformatics, 17,* (2001): 847 - 848.

Zhong, J., Peters, A. H.F.M., Kafer, K., and Braun, R. E. "A Highly Conserved Sequence Essential for Translational Repression of the Protamine 1 Messenger RNA in Murine Spermatids." *Biology of Reproduction,* 64.6, (2001): 1784 - 1789.