PROTEASOMAL GENE DUPLICATIONS AND RECRUITMENT OF TESTIS SPECIFIC EXPRESSION IN DROSOPHILA

by

MEHRAN SOROURIAN

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN BIOLOGY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2008

To my Husband Amin and my Beloved family

ACKNOWLEDGEMENTS

This research has been possible with constant supports and motivations of my supervising professor Dr. Esther Betrán, from whom I learned to be confident and humble toward my goals, which to me is an eternal legacy. I would like to special thank the members of Betán lab whom their support and discussions enhanced my research. Also I would like to acknowledge my academic advisors Dr. Cédric Feschotte and Dr. Shawn Christensen for reviewing my works and their helpful discussions, and invaluable comments that helped me to improve the quality of the research.

I would also like to extend my appreciation to Department of Biology at University of Texas at Arlington, Dr. Esther Betrán and Honors College for providing me the financial support, academic resources, and facilities, all of which made this research possible. I am grateful to all the teachers whom taught me knowledge and wisdom during these years.

Finally, I would like to express my deepest gratitude to my family especially my husband Amin, whom with his endless encouragement and understanding supported my studies.

July 07, 2008

ABSTRACT

PROTEASOMAL GENE DUPLICATIONS AND RECRUITMENT OF TESTIS SPECIFIC EXPRESSION IN DROSOPHILA

Mehran Sorourian, M.S

The University of Texas at Arlington, 2008

Supervising Professor: Esther Betrán

Proteasome is a large multisubunit complex that degrades ubiquitinated proteins in a highly regulated manner in all eukaryotes. Two subcompartments of this subunit are the 19Scap and the 20S core particle with 19S acting as regulatory subunit, it attaches at both ends of the 20Score particle harboring the catalytic domains. In *D. melanogaster* one third of the genes encoding for this subunit were recognized to have many male specific duplicates. In this work, we make use of the new 12 genomes sequences and show that the majority of the genes encoding for the 20S particle gave rise to at least one functional duplicate mainly through retroposition in the past 60 My. Using evolutionary data, we estimated the rate of evolution of 3 retrocopies and compared them to their parental genes. They all evolve faster than the parental genes which could be explained by relaxation of constraint or positive selection. Additional data is needed to discern between these alternatives.

We study the expression of *pros28.1A*, a gene known to have testis specific expression in *D. melanogaster*, in *D. simulans* and *D. yakuba*. In these species we reveled that *pros28.1A* is transcribed in a male specific pattern using different transcription start site. In an effort to understand the motif/s that drive this testis specific expression, we experimentally studied the regulatory region of *pros28.1A* in *D. melanogaster*. Originated through retroposition event via an mRNA intermediate, *pros28.1A* and retrogenes alike usually lack the regulatory sequences. Thus, in order to be transcribed, they must recruit new promoters. Here we show that a short region (i.e. 46bp) very close to the transcription start site of *D. melanogaster* drives the expression of a reporter gene in the testis of transformant flies. However, this region is not likely to drive the testis specific expression of this gene in *D. simulans* and *D. yakuba* which have different transcription start site from that of *D. melanogaster*. This data is in agreement with the previously observed high gene turnover for testis expression in Drosophila.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Gene Duplication and Retroposition

The hereditary information of organisms is carried in DNA: deoxyribonucleic acid. DNA is a complimentary double stranded structure that wraps around each other in right handed fashion. Composed of four nucleotides (Cytosin, Guanin, Thymin, and Adenin) that are bound by phosphodiester bonds, DNA is composed of coding and noncoding sequences. While parts of the coding regions are being translated into proteins, others remain as RNA genes. This is when, noncoding regions (introns, and intergenic regions) often play an important role as regulatory sequences (Lewin 2004). Overall, they all have critical roles in the organism's development, structure and function.

Genomes can consist of few hundred genes in Mycoplasma genitalium (Fraser, Gocayne et al. 1995) to twenty thousand genes in human (2007). This number can even be different between closely related organisms because of gene gains or losses. This change in gene number is an indication of genome evolution and often leads to adaptation. New sequences can originate through different mechanism such as segmental duplications, retroposition (Esnault, Maestre et al. 2000), domestication of proteins from transposable elements (Feschotte and Pritham 2007), *"de novo"* gene formation from noncoding DNA (Levine, Jones et al. 2006), or rearranging the existing ones. Genes can also be lost through mechanisms such as disabling substitutions, and indels (Lewin 2004). For example *Drosophila melanogaster*, an extensively studied eukaryote, has a 170 Mega base (Mb) genome (Ashburner, Golic et al. 2004) that has gained several new sequences in the last few million years. Some examples of newly evolved genes in the *D. melanogaster* lineage are *Dntf2r* (Betran and Long 2003), *mgstl-like-psi* (Toba and Aigaki 2000) and duplications of many proteasomal genes which will be discussed in chapter 2. .

1.1.1. Gene Duplication

Gene duplication is the process by which a gene or a DNA sequence is copied to another place in the genome. At one extreme, one exon can be copied and used in another gene also known as exon

shuffling. At the other extreme, the whole genome can be duplicated (Ohno 1970). Similar to any type of mutation, duplication happens by chance and at random and natural selection selects against or in favor of them (Graur D 1999). If selected in favor, the advantageous mutations could be fixed in the populations thereafter, but when neutral or deleterious, these mutations could either be removed from population or could reach fixation by genetic drift. In the cases where the duplicate copy is fixed in population, there is an additional copy that is free to evolve under different selective constraint than the parental copy. This could lead to neofunctionalization (acquisition of a new function) of one of the copies (Graur D 1999). Besides neofunctionalization, there could be other outcomes after gene duplication. While duplication can be a way of increasing the dose of a gene by making more of the same protein (Ohno, 1970), it can also be a way of fixing heterozygote advantage (Spofford Jul. - Aug., 1969). On the other hand, even if the duplication is neutral it could remain the genome if subfunctionalization occurs. This happens when a gene with expression in several tissues duplicates. In such case, both copies might be preserved if complementary degeneration of their regulatory region occurs resulting in partitioning the expression between the two genes (Lynch and Force 2000). Differentiated gene families are another product of duplication event followed by divergence between the copies (Grisham 1998; Graur D 1999). Thus, duplications are considered to be a major evolutionary force contributing to species diversity and adaptation (Betran, Emerson et al. 2004).

1.1.2. Retroposition

One mechanism of gene duplication is retroposition (Brosius 1999). Retroposition is a process in which new genes are generated at new genomic positions via Target primed reverse transcription (TPRT) (Kazazian 2004) as shown in Figure1.1. This could be mediated by the enzymatic machinery of the Long interspersed elements (LINEs) a type I transposable element. Indeed LINEs can transpose their own transcripts through reverse transcription (cis-effect), but they can also transpose other transcripts back into the genome (trans-effect) (Esnault, Maestre et al. 2000).

The retroposed copies of genes (retrogenes) can be recognized in the genome for their lack of introns, presence of Poly-A tail and target site direct repeats. In *Drosophila*, the last two features are often lost in old retrogenes because of mutation decay (Betran and Long 2002). It has long been recognized

that for retrogenes to lead to functional genes, they must recruit a *"de novo"* regulatory region, carry regulatory regions from the parental gene or insert in front of a region with regulatory capabilities (McCarrey, Geyer et al. 2005; Feschotte 2008). Additionally, TEs could supply "ready to use" regulatory sequences (Feschotte 2008). These sequences (i.e core promoters, enhancers and silencers) which are usually located upstream of the Trascription Start Site (TSS) of a gene, play essential roles in transcription and expression pattern. Part of this work was dedicated to understand the regulatory sequence responsible for transcription of one retroposed copy.

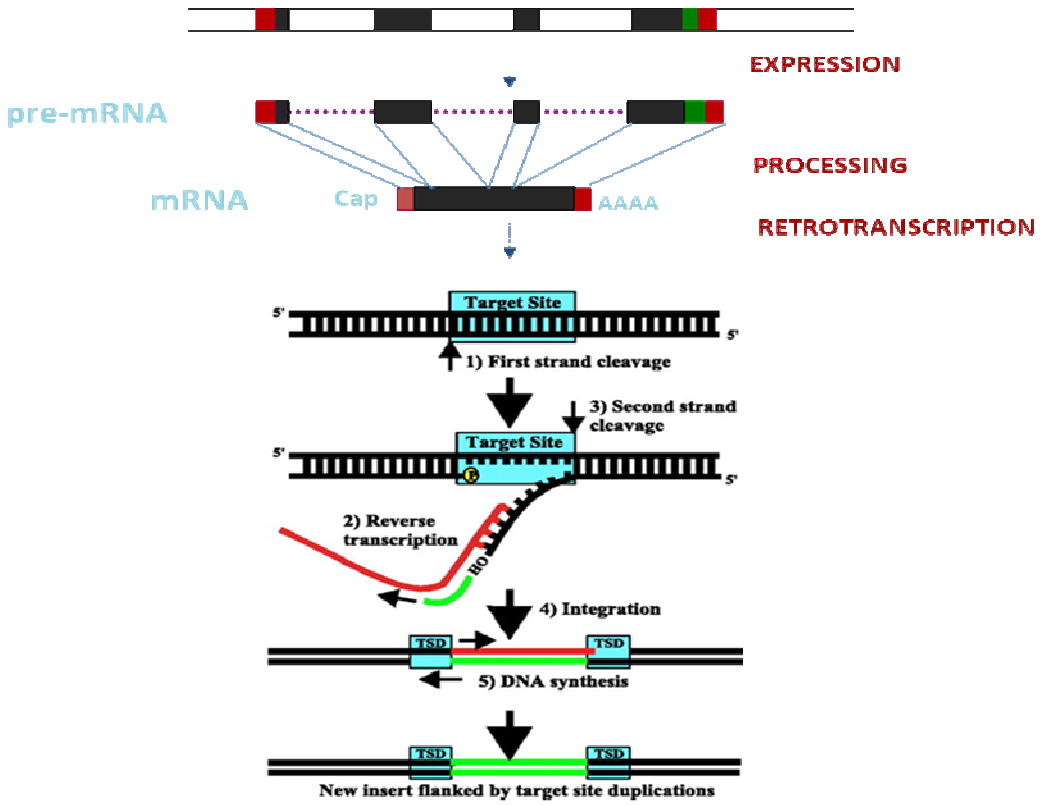

Figure 1.1 Retroposition. Retroposition of a gene into new genomic region by Target Primed Reverse Transcription (TPRT). (figure adapted from (Kazazian 2004)

1.2 Transcription and Regulatory Regions

Understanding the importance of acquiring promoter regions after retrogene formation requires basic knowledge of transcription. Transcription of eukaryotic protein-coding genes is the result of

positioning the RNA Polymerase II in the correct initiation site. When in place, the transcription starts leading to an immature mRNA which by the action of a set of enzymes, gains a 5' cap and polyA tail, and loses its introns. In the cytoplasm, the mature transcript will be translated into a protein (Lewin 2004). What signals the positioning of the RNA Polymerase II are certain DNA regulatory sequences. These specific DNA cis-elements, although diverse, all play a major role in the binding of factors (i.e. transcription factors) that initiate the assembly of the preinitiation complex. Other elements found in cis and trans could work in enhancing or repressing transcription determining the pattern of expression of a gene (Smale and Kadonaga 2003).

1.2.1. Cis-acting Regulatory Elements

Regulating the amount of transcription in a particular cell type is an important task that is accomplished by complex interactions of regulatory proteins with each other and with DNA define a so called cis-acting regulatory region. This region includes the core promoter and other cis-regulatory elements. The most common motifs found in the core promoter region that control the binding of the RNA polymerse II are TATA box, Initiator (Inr), and Downstream promoter element (DPE) as is shown in Figure 1.2. The TATA box is found in 40% of *Drosophila* genes (Ohler, Liao et al. 2002), 32% in human (Smale and Kadonaga 2003) has a sequence of 5'-T A T A (A/T) (A/T) T- 3' sequence that is usually located -25 base pairs (bp) upstream from the transcription start site (Ohler, Liao et al. 2002). The TATA box with the help of initiator element (Inr) attracts the RNA polymerase II to the transcription start site, however due to the poorly conserved nature of Inr, it appears that the TATA box plays a predominent role in directing the polymerase to the site of initiation (Grisham 1998). Inr motif was first identified as a sequence that contains the initiation site and directs the RNA polymerase to the correct site of initiation. However it is known that Inr is also able to redirect the RNA polymerase II to alternative start sites (Smale and Kadonaga 2003). Inr sequence is 5'- T C A(+1) G/T T (C/T) -3' in Drosophila and it is also found in TATA less promoters, DPE less promoters as well as in promoters that contain both motifs (Grisham 1998) .

Another sequence well conserved from Drosophila to human that plays a regulatory role in gene transcription is the DPE. As its name reveals, DPE is located about 30 nucleotides downstream of the initiation site (+1) and has a 5'-G (A/T) C G- 3' sequence. DPE and Inr can function together as a single

core promoter unit where mutations in either of these motifs results in loss of TFIID binding and basal transcription activity (Smale and Kadonaga 2003). The spacing between the two motifs is shown to be sensitive to indels which reduce the transcriptional activity several folds (Kutach and Kadonaga 2000).

Other regulatory elements like promoter proximal elements are located near the core promoter (50-200 bp) and possess one or more binding sites for interaction with DNA-binding regulatory proteins (DBP) to stimulate transcription above basal level. Another group that the DBPs bind to and regulate the mRNA synthesis is the distal enhancer elements. The latter group could be found hundreds of base pairs from the promoter either upstream to the gene or downstream or even in the introns (Grisham 1998; Kanhere and Bansal 2005). Binding of these elements to the DBPs influence transcription through facilitating RNA polymerase interaction with core promoter (Grisham 1998). These elements are needed to establish the pattern of expression of some gene with an exception to the testis specific genes.

1.2.2 The role of promoter in transcription

It has been shown that during transcription, RNA polymerase II and six general transcription factors (GTFs) make up the basal apparatus that bind to the core promoter to initiate transcription. One important factor out of six GTFs (TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) is the TFIID that binds to the TATA box. Recognizing and binding RNA polymerase II to the core promoter leads to the formation of the

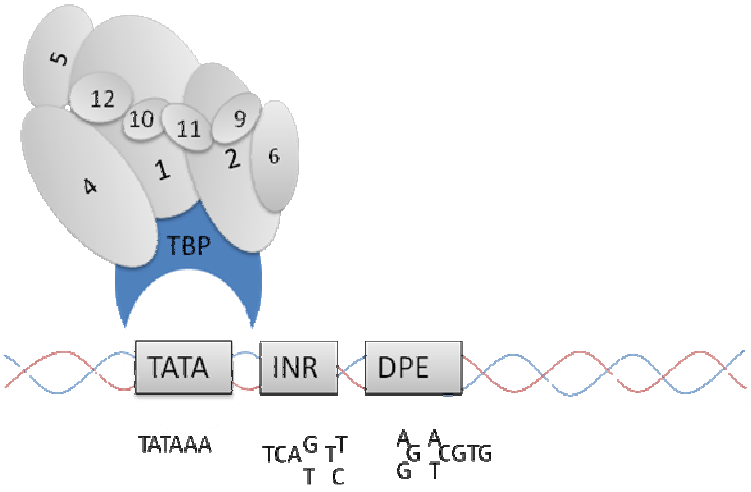

Figure 1.2 TFIID complex in somatic tissues. TFIID recognizes the promoter elements and in association with other TFs initiates transcription. Consensus sequence of each element is shown beneath each element (Redrawn from (Hochheimer and Tjian 2003).

pre-initiation complex. TFIID contains two groups of proteins: the TATA Binding Protein (TBP) which directly recognizes the TATA box and the TBP association factors (TAFs or TAFIIS) that have positive or negative effect on transcription. When the RNA polymerase II, recognizes and binds to the core promoter, it opens up the double stranded DNA. The phosphodiester bonds form between ribonucleotides with the catalytic help of the polymerase in the elongation stage. When the polymerase reaches the termination site, it dissociates thus the transcription stops (Grisham 1998). In the cases of TATA less promoters other promoter elements are recognized by the transcription machinery and initiate the process.

1.2.3. Testis specific TAFs

While TAFs make up major part of TFIID complex (10-14 factors), their role in transcription is not well understood yet. Interaction with transcriptional activators in vitro suggests a role as activators in some promoters. Others interact with DNA motifs and they might have roles in binding of TFIID complex to the promoter element (Hiller, Lin et al. 2001). So far there are five testis specific TAFs known: *cannonball (can), spermatocyte arrest (sa), meiosis I arrest (mia), nonhitter (nht)* and *ryan (rye)* (Chen, Hiller et al. 2005). Expression of these genes such as *can* in primary spermatocytes could be explained by a tissue-specific transcription program that might be needed for differentiation of male gametes (Chen, Hiller et al. 2005). These findings (Chen, Hiller et al. 2005) strongly suggest substitution to a tissue-specific TAF

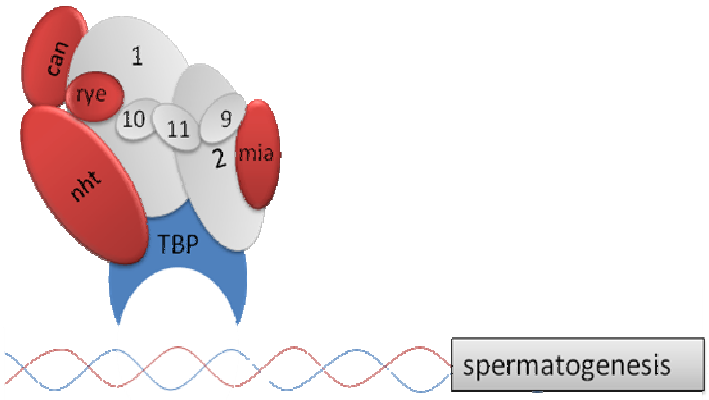

Figure 1.3 TFIID complex during spermatogenesis. Testis specific TAFs replace the ubiquoutous TAFs and initiate transcription of male specific genes in Drosophila. TATA binding protein in blue and the described tTAFs in red when the TAFs with no testis specific homolog are grey (Redrawn from (Hochheimer and Tjian 2003)

isoforms in Drosophila (Hiller, Lin et al. 2001). In such case, the TFIID complex of figure 1.2 is replaced by the one in spermatocytes figure 1.3.

1.2.4 Retrogenes and the importance of acquiring regulatory regions

It is important to note that when a retrogene originates, it will lack most of the regulatory regions except for the downstream elements such as DPE. Lacking these regions could lead to retropseudogene formation and often does (i.e. there are a lot more retropseudogenes in the human genome than retrogenes; (Emerson, Kaessmann et al. 2004) . However there are many instances of retrogenes that are expressed and do so in a particular pattern which is usually different from the parental gene (Yuan, Miller et al. 1996; Ma, Katz et al. 2002; Zhong and Belote 2007; McCarrey Jan., 1994). This could be due to recruitment of a cryptic promoter from the region of insertion or by modifying the pre-existing sequences in the inserted region. Also if the retroposed copy originated from an aberrant transcript that started upstream of the parental TSS, it could carry the parental promoter into the new site. An example of this case has been shown in *Pgk-2* mamalian retrogene. This retrogene formed from an aberrant transcript that is believed to have carried the promoter of the parental gene. Later, the original expression changed and the retrogene has a male specific expression pattern while the parental *Pgk1* expresses ubiquitously (McCarrey Jan., 1994).

1.3 Protein Degradation

There are two types of proteolytic cascades in the cell. The caspase pathway involved in programmed cell death or apoptosis and the ubiquitin-proteasome pathway (Ciechanover and Schwartz 1998). Proteasome is the ubiquitin mediated protein degrading machinery in eukaryotic cells. By selectively destroying the abnormal/mutated and misfolded proteins, it plays important roles in transcriptional regulation as well as regulation of cell differentiation, division and cell cycle (Groll, Ditzel et al. 1997; Ciechanover and Schwartz 1998). The protein degradation occurs via process called ubiquitin protease activity. In this process, the ubiquitin which is an evolutionarily conserved protein of 76 residues is covalently bound to the protein that is selected for destruction. Catalyzed by E1, E2 and E3 enzymes, the ubiquitinated protein is recognized by the 26S proteasome subunit to be degraded (Glickman and Ciechanover, 2002).

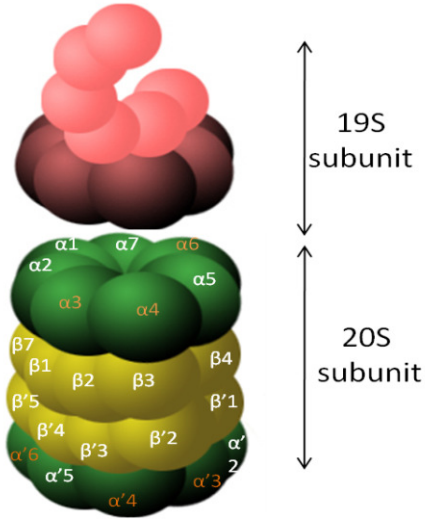

Figure 1.4 Proteasome. Made of two subunits: 19S cap and the 20S subunit. 20S subunit is made of four heptameric ring (α-green) and (β-yellow)

As shown in figure 1.4 proteasome (i.e. the 26S subunit) is made of two smaller subunits: 20S and 19S. Binding at both ends of the 20S, the 19S functions as a regulatory cap while the 20S subunit is a catalytic core. The 18 proteins making the 19S cap are encoded by at least one corresponding gene in eukaryotes (Groll, Bajorek et al. 2000). While the cap plays an important role in recognizing the tagged protein, a ring of 6 proteins at its base with ATPase activity is responsible for unfolding the protein, opening the core particle channel and feeding the protein into the 20S chamber to be degraded. X-ray crystallography studies in yeast and mammalian proteasomes show that the barrel shaped 20S subunit made of 28 proteins is organized as four stacked rings of seven proteins each. This is in an α(1-7)β(1-7)β'(1-7)α'(1-7) ring arrangement and is encoded by at least 14 different genes in eukaryotes (Groll, Ditzel et al. 1997; Groll, Bajorek et al. 2000; Unno, Mizushima et al. 2002). In Archea, the 20S subunit is encoded by a single gene while in eukaryotes each protein is encoded by one corresponding gene (Lowe et al, 1995). Regardless of the gene number encoding for proteasome subunit the structure is conserved from Archea to mammals. While the β rings harbor the catalytic sites, the α rings are responsible in securing the structure, binding to the 19S cap and ensuring the closure of the structure at both ends to

avoid degradation of untagged proteins (Groll, Bajorek et al. 2000; Glickman and Ciechanover 2002; Groll and Huber 2003).

Discoveries in recent years have shown an interesting fact: six out of the twenty eight proteins in the 20S subunit and four of the 19S regulatory cap have gene duplicates encoding for proteins with male-specific functions (Yuan, Miller et al. 1996; Ma, Katz et al. 2002; Zhong and Belote 2007). This is about one third of the genes coding for the proteasome. Such discoveries present us with several questions. Why do we see such high number of duplicates in proteasome? Is there a driving force? Why do we see these genes recruiting male specific pattern of expression? How do they recruit this pattern of expression?

1.3.1 Spermatogenesis

Spermatogenesis in Drosophila is marked by four rounds of mitosis followed by a round of meiosis producing 64 interconnected spermatids (Figure 1.5). Spermatids later go through elongation and individualization producing the condensed motile sperms. It is known that the transcription machinery nearly shuts down at some point during meiosis (Fuller, 1993). Because of this, most of the genes required during later stages of spermatogenesis are transcribed early in meiosis resulting in growth of the primary spermatocytes to 25X of its volume. The 64 interconnected spermatids then individualize and the resulting spermatid undergoes tremendous structural changes to assume the needle shape sperm. During this time, extra cytoplasmic material is degraded, nuclei is condensed and the histones are substituted by proteamines. The resulting sperm then coils up and moves to the seminal vesicle to be stored.

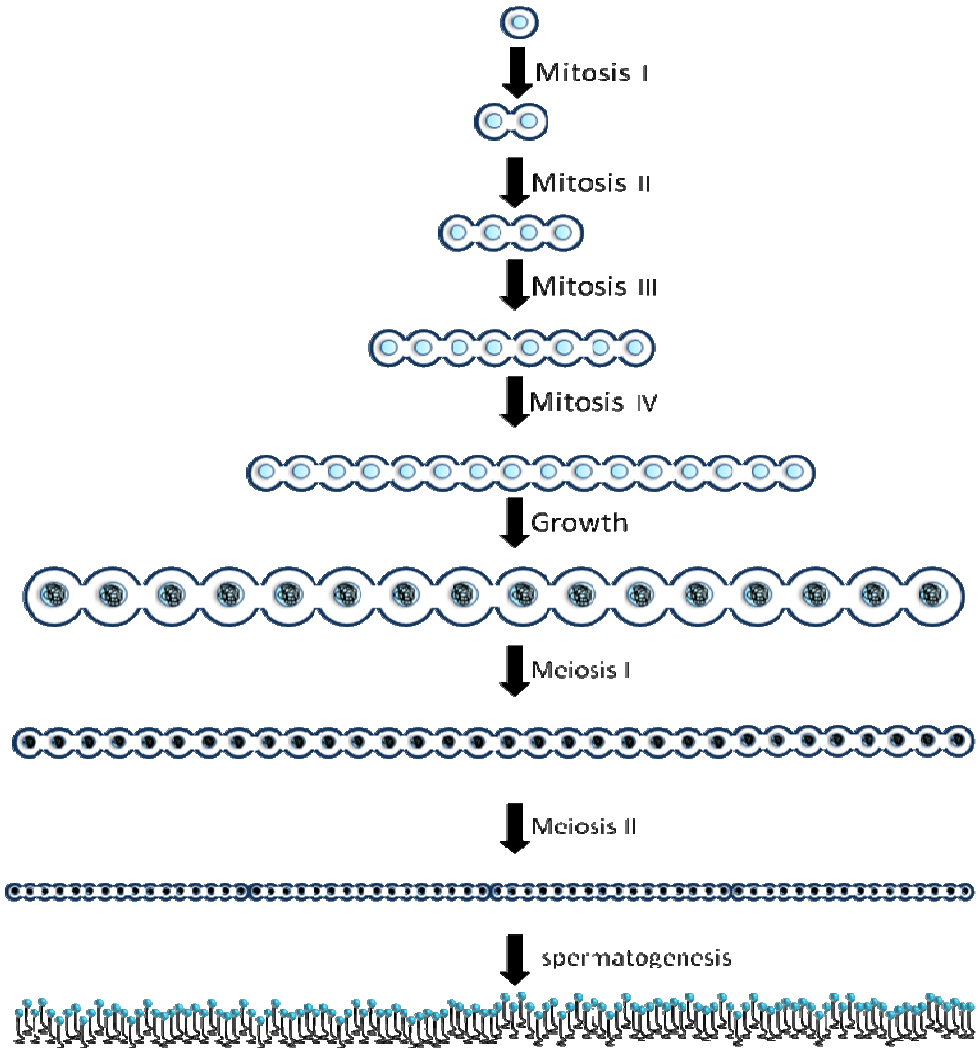

Figure 1.5 Spermatogenesis: 4rounds of mitosis followed by a round of meiosis results in 64 interconnected spermatids that will elongate, and differentiate into mature sperm.

1.3.2 Male specific proteasome duplicates: possible explanations

As we have introduced above, many proteasomal genes have been detected having a testis specific duplicate. In this section, we introduce the possible explanations for this high number of duplications which will help in understanding some of the goals and approaches undertaken in this thesis.

Making “more-of-the-same” could be an explanation for the observed pattern. In the cases where there is a selection for having more of the same protein in a certain tissue, duplication could be an easy way of achieving this. During late stages of spermatogenesis where extensive amount of proteins need to be degraded, overt number of proteasomes are probably needed. Thus, duplication could contribute to cope with this need (Belote and Zhong 2005). In such scenario, the expression of both parental and duplicate copies should overlap during spermatogenesis. In addition, the parental and the duplicate protein should be highly similar.

Another selective explanation for the presence of testis-specific proteasome duplicates could be linked to male germline X inactivation. X chromosome in Drosophila and mouse are heterochromatized to an XY-body during spermatogenesis(Richler, Soreq et al. 1992; McCarrey, Watson et al. 2002). As a result, the genes that are located in the X chromosome are less likely to be accessible for transcription. In the cases where these genes are duplicated to autosomal chromosomes, they could be selected in favor which could describe by the abundance of the X linked parental copies with autosomal duplicates (Betran, Thornton et al. 2002). In this scenario it would be advantageous to have another copy in an autosome that could be transcribed while X is inactive. As result we expect all the duplicates to be X to autosome duplications and keep parental functions.

1.4 Goals of the thesis

As mentioned above, we wonder why there are so many proteasomal duplicates in *Drosophila*? Is there a driving force? Is there an advantage to have a gene copy expressing in the testis? How are these duplicates evolving? What does it take to recruit a male specific promoter? I explored these questions using the following approaches:

- With complete sequence of all 12 Drosophila genomes available (Clark, Eisen et al. 2007), I explored if there are additional/undescribed duplicated proteasome genes and determined their evolutionary history in the Drosophila genus(Chapter 2).
- I studied the mode of evolution of three young (identity to the parental gene higher than 65% amino acid) proteasome duplicates produced through retroposition (Chapter 2).
- I described the pattern of expression and determined the primary transcription start site of the *D. melanogaster* testis-specific retrogene (*prosα4T1*) in *D. simulans* and *D.yakuba* (Chapter 3).
- I determined the minimal upstream sequence of prosα4T1 driving testis-specific transcription of a reporter gene in transgenic flies (Chapter 3).
- Through sequence comparison with parental and with the regions before the insertion of *prosα4T1*, I explore if the promoter element was recruited from the parental gene, from an existing sequence in the region of insertion or from a new transposed sequence (Chapter 3).

CHAPTER 2

PROTEASOME PROTEINS AND THEIR DUPLICATES IN DROSOPHILA

2.1 Overview

In this chapter, with the annotation of 12 recently sequenced Drosophila genomes in hand and the knowledge of previously discovered duplications in *D. melanogaster*, we looked for additional duplicates in other lineages as well as *D. melanogaster*. We describe the structure, age and location of duplicates and explore the mode of evolution of three retroposed copies present in *D. melanogaster*. We also provide information about their pattern of expression when known.

2.2 Proteasomal gene duplications

It was thought that proteasome was a homogeneous protein complex in different tissues. However, recent data obtained in mammals (Gczynclsdkjf et al, 1993), and flies (Yuan, Miller et al. 1996; Belote, Miller et al. 1998; Ma, Katz et al. 2002) shows that some subunits are possibly replaced in some tissues by their counterparts originated through duplication (Belote and Zhong 2005). These duplications might have resulted in subfunctionalization or neofunctionalization. Several studies support subfunctionalization (Ma, Katz et al. 2002; Belote and Zhong 2005; Zhong and Belote 2007) because the genes are expressed complementarily and rescue each others function. However, neofunctionalization and/or specialization of the duplicate cannot be ruled out in other cases. For example in Drosophila, presence of a specialized proteasome functioning during spermatogenesis is highly probable. This is suggested by the fact that 1/3 of the proteasome components have testis-specific counterparts which in some cases function differently from the parental copy (See details below) (Yuan, Miller et al. 1996; Belote, Miller et al. 1998; Ma, Katz et al. 2002). While a tissue-specific, specialized proteasome could be the case in flies and mammals, in plants, the duplication of proteasomal genes is apparently a way to increase the dose and to ensure that sufficient amount of the corresponding protein is available (Yang, Fu et al. 2004; Belote and Zhong 2005).

2.2.1 Duplications in mammals

In mammals, it has been shown that proteasome has an additional cellular responsibility in generating intracellular peptide antigens. When the foreign or mutated peptide is carried to the cell surface, they are recognized by the T lymphocytes which triggers the immune system in destroying the abnormal cell (Gaczynska, Rock et al. 1993; Garcia-Lora, Algarra et al. 2003). Such Immunoproteasomes are clear example of specialized proteasome in mammals. During the immune response, the duplicated form of β1, β2, and β5 which are the catalytic domains of the proteasome are replaced by their counterparts :β1i, β2i, β5i. As a result they form a specialized proteasome that would degrade the proteins in a different fashion (Groettrup, van den Broek et al. 2001).

2.2.2 Duplications in plants

Duplication of proteasomal genes have also been observed in plants but for a different purpose than the immunoproteasomes in mammals. In Arabidopsis, all but 9 genes encoding for the proteasome are encoded by two genes with their encoded protein detected in purified proteasomes (Yang, Fu et al. 2004). When compared to their parental genes, these duplicate copies have same pattern of expression and are at least 88% identical to their parental genes. This high degree of similarity along with data showing no phenotypic affect in mutants lacking one of the duplicates could indicate that these duplicates are functionally redundant and that they serve to ensure sufficient amount of corresponding protein (Belote and Zhong 2005).

2.2.3 Duplications in Drosophila

As commented above, studies investigating the duplicates of proteasome in Drosophila shows over one third of the genes encoding the 20S subunit are duplicated (Yuan, Miller et al. 1996; Ma, Katz et al. 2002; Zhong and Belote 2007.). These duplicates have shown to be often differentially expressed compared to the parental genes and it has been experimentally shown in some cases that the loss of the duplicate leads to sterility. For instance Rpt3 a protein that is part of the 19S regulatory lid, has a duplicated copy, *Rpt3R* which caused male sterility when knocked out (Belote and Zhong 2005) This has also been observed with knockouts of some duplicates involved in the formation of the 20S base. A

knockout of *prosα6T* that is the retroduplicate of *prosα6* leads to sterility in males. The parental gene is shown to rescue this phenotype when expressed during the late stages of spermatogenesis, a stage at which the *prosα6T* transcript is known to be transcribed (Zhong and Belote 2007). Although this is an indication of similarity in function, other genes like *prosα3* and its male specific retrocopy (*prosα3T*) show different effects when over-expressed, suggesting functional divergence. Indeed, when *prosα3T* is ectopically over-expressed in soma it leads to pupal lethality while the over-expression of the parental copy in the same tissues has no lethal effects (Belote and Zhong 2005).

2.3 Recurrent duplication in Drosophila

Seeing vast numbers of proteasomal gene duplications in *D. melanogaster*, we were motivated to screen for more duplicates with the latest annotations of all 12 sequenced genomes of Drosophila available. This was also done to reveal if some of the same genes have been duplicated in other lineages.

2.3.1 Searching the genome for duplicates

Blast searches (tblastn) of protein coding sequence of all 32 genes encoding the 26S subunit against all 12 annotated genomes (flybase) revealed more duplicated isoforms than previously described. These duplicated copies were first checked for their identity using alignments with the parental copies and then checked for presence and absence of annotated introns to determine the mechanism of duplication. It has been observed that the occurrences of insertions of sequences (i.e transposable elements) or disablements (i.e. short indels or STOP codons) into the transcripts are often mistakenly annotated as introns. To avoid such mistakes, alignment of raw sequence with the parental copy were performed and the position of the corresponding introns were compared with that of the parental gene. In 80% of the cases introns were absent and since absence of introns is a classic way of detecting retroposed duplicates, the sequences which lacked introns were considered as retrocopies.

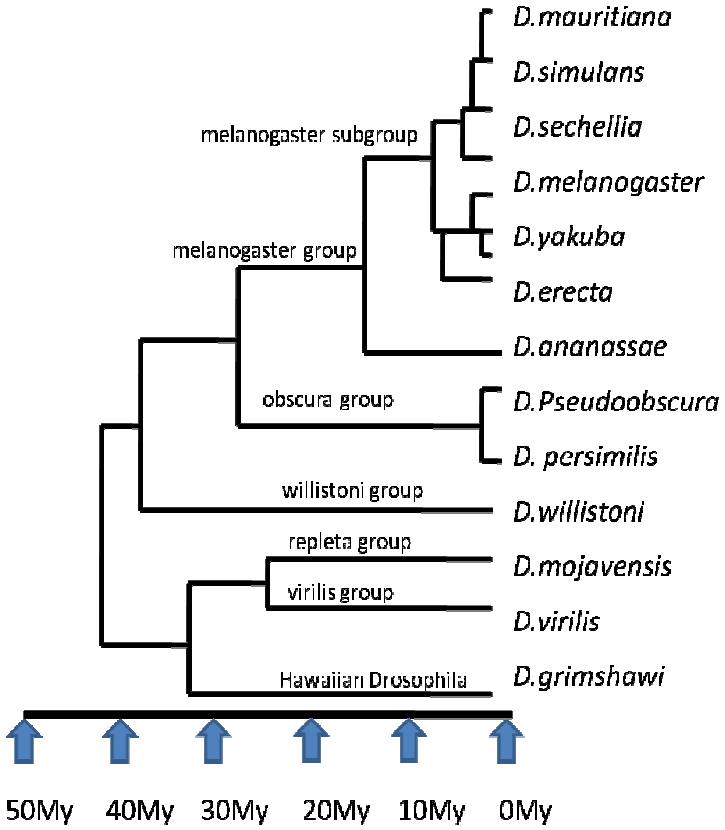

Figure 2.1 The phylogenetic tree of 12 sequenced genomes of Drosophila. Generated and revised from flybase (Wilson, Goodman et al. 2008)

The sequences retrieved were then checked for possible functionality. In the cases were sequence of the transcript were not available the open reading frame was determined by ORF finder (NCBI). The translated sequences were then aligned to the parental translated sequence for similarity and possible functionality.

Many duplicates such as retroduplicates lack regulatory sequences, if landed in the region not capable of recruiting the transcription machinery, these duplicates are usually doomed to be pseudogenized. To check if the duplicate isoforms were a pseudogenized copies or possibly functional, their translated sequence were checked against premature stop codons, large insertions or deletions which normally affect the protein functionality. In the cases of few amino acid insertions, the protein was

reported as functional and its translated sequence was aligned with the parental amino acid sequence for similarity.

Dating these duplicates was performed by comparative sequence analysis where the duplicated locus was screened in other species for presence/absence. We also looked at the syntenic regions for any traces of the pseudogenized copies that might have not been detected by the initial blast. Additionally, a phylogenetic reconstruction approach was taken where the translated protein sequence of the newly found duplicates was aligned with the translated sequence of parental and other duplicated copies (if applicable) using Clustal W (Larkin, Blackshields et al. 2007). Genes found in different lineage that are not in syntheny are considered to be orthologous if they do not cluster together in the phylogeny. For example the tree generated from alignment of the *prosα4* (the parental copy) and its three independent duplicate copies (Figure2.2), shows that *pros28.1A* and *pros28.1B* which were previously described (Yuan, Miller et al. 1996) were independently duplicated. And besides the annotated *pros28.1B* that is present in 12 sequenced genomes, there is an additional duplicate copy that is annotated as *pros28.1B* in *D. virilis*. In the X chromosome this is an independent duplication that based on homology is most likely derived from the *pros28.1B*. To differentiate these two duplicates the one in *D. virilis* is shown as *"pros28.1B".* Another duplicate of *prosα4* occurred in the ancestor of obscura group and is annotated as GA25292 in *D. pseudoobscura* and GL26115 in *D. persimilis*.

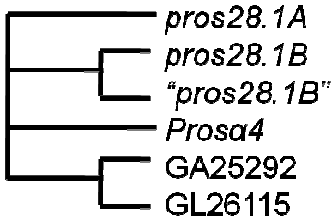

Figure 2.2 Recognition of orthologous sequences. An example of the generated tree showing clustering of orthologous duplicates with the parental gene (*prosα4)* and within duplicates

2.3.2. Proteasome protein duplications

Previously, 5 out of 14 genes encoding for the 20S subunit of proteasome complex were found to have duplicates (Yuan, Miller et al. 1996; Ma, Katz et al. 2002). Here we show that all the genes encoding

for the 20S except α2, α5, and β3 have duplicates in at least one lineage and they are mainly generated through retroposition (Table 2.1). The 11 newly discovered duplicate copies in the base along with the 3 in the lid makes up 14 novel duplicates 10 of which are identified as retroposed copies. These duplicated copies have from as high as 81% amino acid identity with their parental counterpart to as low as 43% identity. This variation in percent identity not only could pertain to the age of these duplicates but also in some cases could point to functional divergence. For instance, while duplication of both *prosβ2R2* and *prosα4T1* dates back to between 8-14Mya according to their phylogenetic distribution, the amino acid identity between the duplicates and their parental copies is drastically different (35% compared to 74% respectively). Since the transcripts of both duplicate copies were detected in males of *D. melanogaster*, the difference is not due to pseudogenization of the copy. This might indicate functional divergence between the parental protein compared to the duplicate protein in the case of *prosβ2R2* while the *prosα4T1* might have maintained main function of its parental gene. We also have to keep in mind the possibility that *prosβ2R2* could be older than inferred and it might not have been detected due to loss, gaps in the assemblies or divergence and reorganizations.

Microarray data along with northern blot studies in *D. melanogaster* show that there is a difference in transcription pattern of duplicated genes compared to the parental copies. This bias is without exception bound to expression in male germline with no expression in females. With presence of a testis specific proteasome and the information on male specific expression of all duplicate copies, it is very likely that the additional duplicated forms that we described here are as well expressed in a male specific fashion. Of course the expression pattern of these newly discovered duplicated copies remains to be investigated in future.

Table 2.1 proteasomal duplicates†

house keeping	Map	additional isoform	Map	Type	Detected in species	protein identity	Male specific
prosα1	2R	GJ15627	X	Retrogene	virlis	57%	
		GL12054	3R	Retorgene	persimilis	51%	
prosα2	3R	N/A					
prosα3	2R	prosα3T	3R	Retrogene	melanogaster subgroup	58%	male specific
		GE14482	2R	Non-functional	yakuba		
		upstream CG17652	2L	Non-functional	Willistoni		
		upstream GK20199	2L	Non-functional	Willistoni		
		GL22394,GA28930	2L	Retrogene	obscura group	57%	
prosα4	X	GL26115,GA25292	2L	Duplicate	obscura group	81%	
		GL21346	X	Non-functional	persimilis	86%	
		pros28.1B(pros$_{\alpha 4T2}$)	2R	Duplicate	all 12 genome	60%	male specific
		"pros28.1B"	X	Duplicate	virilis	51%	
		pros28.1A (prosα4T2)	3R	Retrogene	melanogaster subgroup	74%	male specific
prosα5	2R	N/A					
prosα6	2L	prosα6T	2L	Retrogene	melangoaster and obscura group	64%	male specific
		GJ15897,GI17321	3R	Retrogene	virlis and mojavensis	53%	
prosα7	2R	GM20573	2R	Non-functional	sechellia	53%	
		GF14761	2L	Retrogene	annassae	77%	
prosβ1	2R	GL26231,GA28089	2L	Retrogene	obscura group	53%	
		intron of GL25556, GA28033	2L	Retrogene	obscura group	53%	
prosβ2	3L	prosβ2R1	X	Duplicate	all 12 genomes	62%	male specific
		prosβ2R2	3R	Duplicate	melanogaster subgroup	35%	male specific
Prosβ3	3R	N/A					
prosβ4	2L	prosβ4R1	2L	Duplicate	melanogaster subgroup		male specific
		prosβ4R2	2L	Duplicate	all 12 genomes		male specific
prosβ5	2R	Prosβ5R1	2R	Duplicate	all 12 genomes	53%	male specific
		Prosβ5R2	2L	Retrogene	melanogaster subgroup	47%	male specific
prosβ6	3L	Gj15881	X	Retrogene	virlis	43%	
		Gj1596	X	Duplicate	virlis	43%	
prosβ7	3R	GF3325	3R	Retrogene	annassae	64%	
Rpt1	2R	GK10072	X	Duplicate	willistoni	99%	
Rpt2	3R	N/A					
Rpt3	X	Rpt3R	3R	Duplicate	all 12 genomes	77%	male specific
Rpt4	X	Rpt4R	3L	Retrogene	all 12 genomes	82%	male specific
Rpt5	3R	N/A					

Table 2.1-continued

house keeping	Map	additional isoform	Map	Duplicate	Detected in species	protein identity	Male specific
Rpt6	X	CG2241	3R	Duplicate	all 12 genomes	88%	male specific
Rpn1	3L	N/A					
Rpn2	3R	N/A					
Rpn3	2L	N/A					
Rpn5	3R	N/A					
Rpn6	2R	N/A					
Rpn7	3R	N/A					
Rpn8	3R	N/A					
Rpn9	2R	N/A					
Rpn10	3L	GK25574	2L	Retrogene	willistoni	84%	
Rpn11	2L	N/A					
Rpn12	3L	CG11552	3L	Duplicate	all 12 genomes	49%	male specific
Uchp37	3L	CG1950	X	Retrogene	Sophophora	56%	male specific
		GL20225,GA26899	X	Retrogene	obscura group	55%	

†Duplicate subunits making the 26Sproteasome with genes encoding for α subunits in green, the β subunits in orange and the 19S cap in pink are shown. Genes that were previously characterized have data supporting their male specificity. Additional isoforms are represented by their annotated name (CG: *D. melanogaster*, GE: *D. yakuba,* GF: *D. annassae*, GL: *D. persimilis,* GA: *D. pseudoobscura,* GK: *D. willistoni*, GI; *D. mojavensis*, and Gj: *D. virilis*,) or location of the gene. N/A indicates absence of detectable duplicates. Retrogenes were characterized by absence of introns. Presence is reported when gene is found but it could be actually older (see identity values).

2.3.3. Discussion

From the 13 proteasomal duplicates present in *D. melanogaster*, 8 originated more than 50Mya. According to the phylogenetic footprinting, these duplicates of *α4T2, β2R1, β4R2, β5R1, Rpt3R, Rpt4R, Rpt6R*, and *Rpn12R* are present in all 12 sequenced Drosophila genomes. Interaction map of 26S proteasome in *C. elegans* shows that *Rpt4* not only interacts with α4, but also with *Rpt3, Rpt6* and *Rpt5* based on yeast two hybrid system (Anne Davy 2001; Coux 2003).If the duplicates of these genes interact as well in Drosophila, then it is possible that duplication with male bias expression of one, was the selective force in fixing some of the following duplicates to form a specialized proteasome. Also it is interesting to note that from these 8 duplicates four are X to autosomal duplicates: *α4RT2, Rpt3R, Rpt4R,*and *Rpt6R*. This could support another hypothesis put forward recently, i.e. that the movement out of X initiated the male specific proteasome basically to avoid the X inactivation that is known to occur

during spermatogenesis (Betran, Thornton et al. 2002). From the non X to autosomal duplicates, it is also remarkable to note that the β2 and β5 which are the main catalytic domains involved in protein cleavage gave rise to duplicate copies with male specific expression: *prosβ2R1* and *prosβ5R1* respectively. This might indicate that the specialized proteasome that began to form more than 50 Mya was possibly a specialized proteasome with different catalytic sites capable of cleaving the proteins in the sperm in a different pattern from that of the somatic proteasome.

It is possible that after fixation of initial duplications, either by chance or because of the male X inactivation process selection favored more duplicate copies thereafter resulting in further specialization of the proteasome (see figure 2.3). Interestingly, under this hypothesis, there would be selective advantage of fixing additional duplications in every lineage and consistent with this we found that some genes have duplicated recurrently. For example in the case of *prosα4*, aside from the male specific duplicate *prosα4T2* detected by Yuan, and Miller present (Yuan, Miller et al. 1996) in all 12 genomes, another duplicate copy *prosα4T1* arose 8-13Mya. This pattern is widely seen in many of the proteasomal duplicates: *prosβ2, prosβ4*, and *prosβ5*. This is in contrast with others like prosα7, *prosβ7* that are duplicated in a single lineage like *D. ananassae*. Also others like *prosα3* have a duplicate present in all species of melanogaster subgroup and one in species of obscura group. *Prosα6* also shows similar pattern to *prosα3*. *Prosα6* has given rise to two independent duplications one in species of melanogaster and obscura group and one in *D. virlis* and *D. mojavensis* lineage almost covering all 12 sequenced genomes.

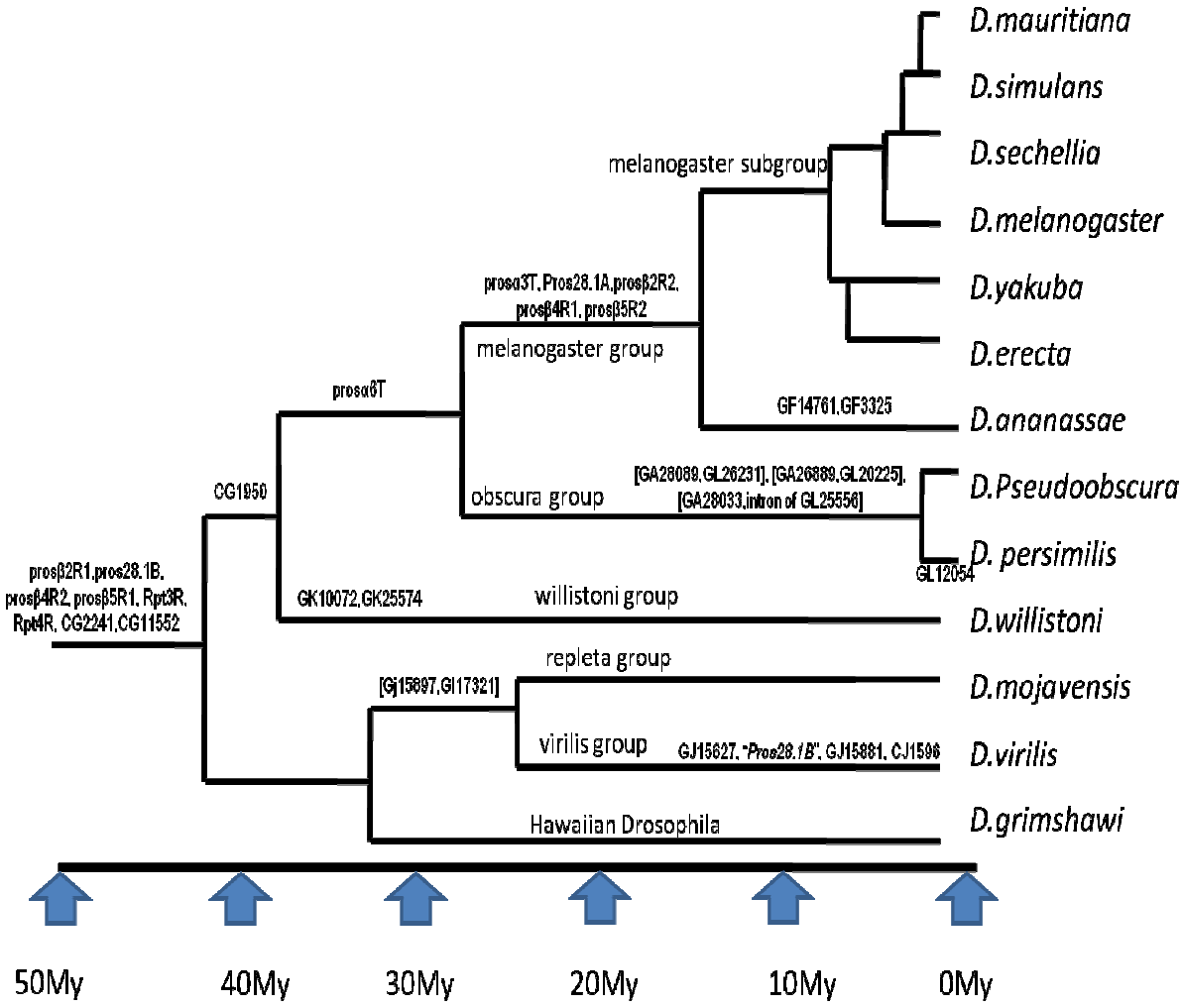

Figure 2.3 Phylogenetic tree with functional duplicates.

To investigate if male germ line specific duplicates lead to a more specialized proteasome, we explore the mode of evolution of some male specific duplicates (see below). In the case of “more-of-the same” hypothesis, we expect parental and derived genes to be very similar and evolve at comparable rates. In subfunctionalization events, pattern of expression of genes should be complementary and, again, we expect most parental and their duplicates to evolve at similar rates and retain high level of identity. However if after subfunctionalization specialization occurs, the genes might be more divergent and possibly evolve at different rates. For example, we might see that the duplicated protein sequence changed rapidly after duplication. Duplicate genes might also be under different selective pressure

(higher constraint, positive selection or relaxation of constraints). Analysis of the mode of evolution should however be complemented with studies of gene function (i.e mutant analysis), rescue experiments and ectopic expression experiments to get an insight into the function of these duplicates.

2.4 Mode of evolution of three male germline proteasome retrogenes

Using Phylogenetic Analysis based on Maximum Likelihood (PAML) software (Yang 1997) some duplicate copies mode of evolution were investigated in the retrogene compared to the parental gene. The program estimates values of ω=dN/dS (nonsynonymous substitutions per nonsynonymous site over synonymous substitutions per synonymous site) under a particular branch model using maximum likelihood. The comparison of different branch models likelihood can then reveal the different modes of evolution in particular branches. This is a way of testing for the nature and intensity of selection and the dN/dS gives an idea of the type and strength of selection acting on a particular branch. A protein is said to be under purifying selection when the dN<dS. In this case, there is a selection against mutations changing the amino acid composition of protein. A protein is said to be evolving under positive selection if the dN>dS value and the ω>1. In this scenario, there is a selection on nonsynonymous substitutions that are likely to change the protein function. If a protein is evolving neutrally this ratio is either equal to one or is close to one. We should keep in mind that if a functional gene (i.e. a gene that is transcribed and shown protein constraint at the population level) is evolving with a rate ~1, it is possible that it is under positive selection as well.

The branch models allow ω to vary among branches or sets of branches and to make comparisons to assess statistically whether rates differ between branches. We use several branch models: one ratio model, two ratio model and three ratio model. The one ratio model sets the ω to a single value for all branches. In the two ratio models the software calculates two ω values for two sets of branches (i.e. parental gene lineages [ωp] and retrogene lineages [ωr]; Figure 2b). In the three ratio model, three different rates are estimated (i.e. parental gene lineages [ωp], lineage after duplication [ωd] and other retrogene lineages [ωr]; Figure 2c). To detect selection in the retrogene, an additional model was run. In this model ωr was fixed to 1 and its log likelihood value was compared to the log likelihood value of the model that estimates ωr. If the original value is significantly better from the one with ωr=1, we

infer positive selection the ωr is >1 and purifying selection if ωr is <1. A tree is provided for all this comparisons. These models were compared by calculating two times the log likelihood values and comparing to a χ^2 distribution with degrees of freedom equaling the difference in number of parameters estimated by each model.

We performed these analyses for three young (i.e. < 35 my old) male germline specific retrogenes present in *D. melanogaster*: *prosα3T, prosα4T1* and *prosα6T.*

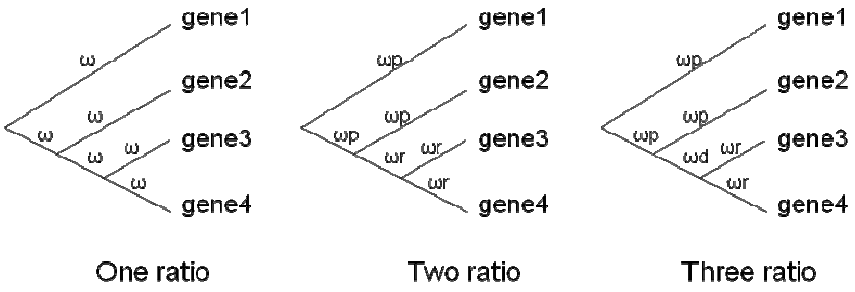

Figure 2.4 Three models used in PAML analysis

2.4.1 Prosα3T

With 58% amino acid identity with the parental copy, *prosα3T* was generated through a retroposition event 8-14 Mya. Northern blot studies in addition to the transgenic studies (Ma, Katz et al. in 2002) characterized this duplicate copy to be male and testis specific expressed after meiosis II in the 64 spermatids stage as well as later during elongation and individualization and in mature sperm, and always in the nucleus. The parental gene *prosα3* is expressed abundantly throughout testis in spermatogonia, primary spermatocytes, secondary spermatocytes, 64 haploid spermatids, and elongating spermatids (Ma, Katz et al. 2002).

Table 2.2. shows the results of our PAML analyses. The comparison between the one ratio model

Table 2.2 The PAML results for *prosα3T*

Model	Log likelihood	Parameter estimates
One ratio	-4399.72	ω=0.08582
Two ratio	-4348.32	ωp=0.0163, ωr=0.221
Two ratio	-4401.99	ωp=0.0168, ωr=1
Three ratio	-4348.27	ωp=0.0163, ωr=0.2228, ωd=0.1755

and the two ratio model shows that the two ratio model is significantly more likely than the one ratio model(X^2=102.8; d.f.=1 ; P<<<0.01) and the three ratio model(X^2=0.0922; d.f.=1 ; P=0.75). This reveals that the retrogene evolves differently from the parental gene (~13X faster with ωr:0.221 relative to ωp:0.0163) and this rate was not different immediately after duplication. The relatively low ωr value (0.221) in *prosα3T* is an indication of purifying selection. In order to obtain statistical support of purifying selection, we compared the two ratio model to a two ratio model with the retrogene's ω fixed to 1. With a significantly lower log likelihood value (X^2=107.36; d.f.=1 ; P<<<0.001) compared to the two ratio model, the retrogene was shown to be under purifying selection. Data supporting purifying selection coupled with fast evolution of *prosα3T* indicates overall purifying selection with either positive selection or relaxation of constraint in some sites of the retrogene. In agreement with the different rate of evolution between parental gene and retrogenes, the over expression of *prosα3T* in larvae leads to pupae lethality when over expression of *prosα3* has no effect on pupae viability (Belote and Zhong 2005) consistent with a difference in function of the two proteins.

2.4.2. Prosα4T1

With 74% amino acid similarity to the parental copy, *prosα4T1* was generated through retroposition event 8-14 Mya similar to that of *prosα3T*. It is only present in species of the melanogaster subgroup. Northern blot studies in addition to the transgenic studies of Yuan, Miller et al. in 1996 characterized this duplicate copy to be male and testis specific expressing during late stages of spermatogenesis with spermatogonial cells deficient of its transcript. In contrast, the parental gene *prosα4* is expressed abundantly in male and females tissues of *D. melanogaster* (Yuan, Miller et al. 1996).

Prosα4T1 also has shown to fit the two ratio model significantly better than the one ratio model or

Table 2.3 The PAML results for *prosα4T1*

Model	Log likelihood	Parameter estimates
One ratio	-3478.66	ω=0.08532
Two ratio	-3440.49	ωp=0.0315, ωr=0.2181
Two ratio	-3483.03	ωp=0.0236, ωr=1
Three ratio	-3439.41	ωp=0.0325, ωr=0.2468, ωd=0.1509

the three ratio model (X^2=76.34; d.f.=1 ; P<<<0.001 and X^2=2.173; d.f.=1 ; P=0.1) and it is also clearly under purifying selection with a significantly lower log likelihood value (X^2=85.08; d.f.=1 ; P<<<0.001). This result is contradictory to the analysis of Torgerson and Singh (2004), who inferred that the rate of evolution increased after duplication (Torgerson and Singh 2004). However, we noted that the tree under which the models were calculated was different and probably inaccurate. Looking at the mode of evolution of this gene compared to its parental gene, *prosα4T1* evolves ~7X faster than its parental copy (0.2181 compared with 0.0315) suggesting either positive selection or relaxation of constraint in some sites of the retrogene. While Torgerson and Singh (2004) found high levels of polymorphism in this gene which would point to relaxation of constraint, functional analysis similar to the ones performed on *prosα3T* remain to be performed in order to assess if there are functional differences between the parental and retrogene copy.

2.4.3. Prosα6T

With 64% amino acid identity with the parental copy, *prosα6T* was generated through a retroposition event 25-35 Mya. This retroposed copy is present in the species of melanogaster group as well as the obscura group. Transgenic studies by Zhong and Belote in 2007 showed that this gene is expressed during meiosis, and spermatid differentiation and individualization. Using florescence reporter gene assay, expression was detected prominently in nuclei & cytoplasm of mature sperms while the parental copy was shown to be ubiquitously expressed in testis but fading away after meiosis (Zhong and Belote 2007).

Table 2.4 shows the results of the PAML analyses of *prosα6T* and its parental. We observe that the one ratio model is significantly less likely than the two ratio model (X^2=4.364; d.f.=1 ;P<0.05).

Table 2.4 The PAML results for *prosα6T*

Model	Log likelihood	Parameter estimates
One ratio	-6098.56	ω=0.1413
Two ratio	-6096.38	ωp=0.1201, ωr=0.1642
Two ratio	-6270.12	ωp=0.107, ωr=1
Three ratio	-6095.78	ωp=0.1189, ωr=0.1573, ωd=0.2313

Retrogene is evolving 1.4 times faster than the parental gene (ωp=0.1201, ωr=0.1642). The two ratio model was not significantly different from the three ratio model revealing no change in rate of evolution of the retrogene after duplication (X^2=1.203; d.f.=1 ;P=0.25). *Prosα6T* shows clear signs of being under purifying selection as well. When we compare the two ratio model with the two ratio model with the retrogene's ω value fixed to 1, the difference is significant (X^2=347.5; d.f.=1 ; P<<<0.001). The results support that the retrogene is evolving faster than the parental and either positive selection or relaxation of constraint could explain this observation. Polymorphism data might reveal the actual reason for the faster evolution. However experimental data seems to indicate that they have similar function (Zhong and Belote 2007). While the knock out mutants of *prosα6T* are shown to be male sterile with disruption of actin cone movement during sperm individualization, the *prosα6* (parental gene) is able to rescue the phenotype. Although there is no indication of the level of rescue, but it supports the fact that the two copies are functionally redundant. Thus, this could be a case where subfunctionalization have occurred.

2.5 Concluding remarks

In Drosophila, it has been observed that the duplicated proteasome genes often become partitioned into those encoding subunits that are expressed through development and those that are limited to the late stages of spermatogenesis. We provide data that reveals that in other lineages additional duplicates remain to be studied for the same pattern of expression. The driving force could be subfunctionalization, neofunctionalization or more of the same. Different duplicates might fit different possibilities. *Prosα6T* could be an example of subfunctionalization: expression pattern of *prosα6* and *prosα6T* does not overlap, the loss of *prosα6T* leads to infertility but can be rescued by the parental gene and the new gene is evolving only 1.4 faster than the parental gene. *Prosα3T* could be an example of neofunctionalization: expression pattern of *prosα3* and *prosα3T* overlaps, overexpression of *prosα3T* in larvae leads to lethality and the new gene is evolving 13 times faster than the parental gene. Additional mutational, ectopic expression and rescue studies need to be carried out for other genes. The rate of evolution data on *prosα4T1* is compatible with the three possibilities despite the fact that it is evolving 7 times faster than the parental. We do not know if the expression pattern of parental gene overlaps *prosα4T1*.

CHAPTER 3

PROS28.1A RECRUITED A SHORT PROMOTER FOR TESTIS SPECIFIC EXPRESSION

3.1 Overview

In this chapter *pros28.1A* (*prosα4T1*) a young retroduplicate copy of the *pros28.1* (*prosα4*) is shown to recruit a short *"de novo"* male specific promoter located upstream of close to the transcription start site of *Drosophila melanogaster*. Due to the conserved expression pattern among the species harboring this insertion, it was initially expected that a promoter common to those species was recruited prior to their speciation. However the use of different transcription start site in *D. simulans* and *D. yakuba* might indicate that either these species use the same promoter as a downstream element or they have recruited another one Which is we are currently exploring.

3.2 Introduction

One of the pathways in degrading the unwanted proteins is mediated by attachment of ubiquitin to the targeted protein and its destruction by the proteasome. This highly regulated process is implicated in cell differentiation, transcriptional regulation and cell cycle progression as well as ridding cells of the mutated proteins (Glickman and Ciechanover 2002). One of the subcomponents making this highly conserved structure is a barrel shape core particle (20S subunit). Encoded by 14 genes (7 α and 7 β), it forms 4 heptameric rings in α(1-7)β(1-7)β'(1-7)α'(1-7) pattern which plays main role in degrading proteins into small polypeptides that the proteases will further break into individual amino acids. In Drosophila, out of the 14 genes encoding the core particle, 6 have duplicates showing male specific expression (Yuan, Miller et al. 1996; Ma, Katz et al. 2002; Zhong and Belote 2007). Expression and functional data support the existence of a testis specific proteasome in Drosophila.

α4 subunit one of the core particle components is encoded by *pros28.1,* a gene located on the X chromosome of *D. melanogaster.* With two introns, it encodes for 248 amino acid long protein that is expressed ubiquitously in *D. melanogaster* (Hass C 1990). While conservation through evolution indicates its important 'housekeeping' role in the proteasome, *pros28.1* has given rise to two duplicates both of

which are in autosomal chromosomes and have a male specific pattern of expression. This is another example of X to autosomal movement of genes in Drosophila possibly to avoidfor due to the X inactivation during spermatogenesis (Betran, Thornton et al. 2002; Emerson, Kaessmann et al. 2004). From the two duplicates of *Pros28.1, Pros28.1B* is an older non-retroposed duplicate (Yuan, Miller et al. 1996; Belote, Miller et al. 1998; Betran and Long 2002) shown to be transcribed during spermatid elongation (Yuan, Miller et al. 1996) while the *Pros28.1A,,* the younger retroposed copy, is transcribed in primary spermatocytes as well as spermatid elongation stage (Yuan, Miller et al. 1996). *Pros28.1A* is retroposed into the 3rd intron of *CG31203,* which is a gene with an unknown function (Wilson, Goodman et al. 2008). The presence of *pros28.1A* in the melanogaster subgroup and its absence in the other lineages of Drosophila suggests that the retroposition event occurred 8-13Mya (Bai, Casola et al. 2007). Retroposition is a type of gene duplication when a new gene is generated in a new genomic position via reverse transcription of an mRNA intermediate (Betran, Thornton et al. 2002). This reaction is likely catalyzed by RT of LINE-like transposable elements which mistakenly act on a "host" gene transcript inserting the resulting cDNA into the genome via Target Primed Reverse Transcription process (Esnault, Maestre et al. 2000; Kazazian 2004). Hallmarks of these sequences are lack of introns, presence of Poly-A tail and target site direct repeats. In *Drosophila*, the last two features are often lost in old retrogenes (Betran, Thornton et al. 2002). It has long been recognized that for retroposed copies of genes to lead to functional genes, they must recruit a "de novo" regulatory region, carry regulatory regions from the parental gene or insert in front of a region with regulatory capabilities (McCarrey Jan., 1994). It has also been suggested that new genes could recruit pattern of expression from the surrounding chromatin context (Kalmykova, Nurminsky et al. 2005), although little evidence is supporting any of these possibilities as a general mechanism.

In this work we sought to identify and characterize the promoter driving the testis- specific expression of *pros28.1A* using reporter gene in transgenic flies. Previously it was suggested that male specific promoters have high turnover in Drosophila genome (Zhang, Sturgill et al. 2007). Here we present a case where a retroposed gene *(pros28.1A)* has recruited a short male specific promoter close to the TSS.

3.3 Material and Methods

3.3.1 Strains used

D. melanogaster (Besançon; P. Gilbert), *D. simulans* (Florida; J.coyne), *D. mauritiana* (72;), *D. santomea* (10; fromSantome and Lachaise 1998), *D.teisseri* (118.2; Lemeunier and Ashburner 1976), *D. yakuba* (115; Lemeunier and Ashburner 1976), and *D.erecta* (154.1; Lemeunier and Ashburner 1976). The strains were grown in standard corn media at 25ºC.

3.2.2 DNA samples and sequencing

Sequence of the CG31203 intron along with pros28.1A coding sequence and its flanking sequences was obtained from FlyBase (Wilson, Goodman et al. 2008). In species that this sequence was not available (i.e D. santomea, D. teisseri, and D. mauritiana), genomic DNA was extracted from single female fly using the Puregene kit. Using the Oligoprimers 5´TTAGGGTTCGGCTTTCCGTA3´, 5´ACCTGCTATCCTGGGTGATC3´ and 5´CAACGCTATCCTGTGTCGC3´ ordered to Integrated DNA Technologies Inc. Pros28.1A and its flanking sequence was PCR amplified in D. mauritiana, D. teissieri and D. santomea. PCR products were then sequenced directly after purification (Qiagen kit) on an ABI automated DNA sequencer (Applied Biosystems) using fluorescent DyeDeoxy terminator reagents. Sequences were obtained from both strands to confirm every position, contigs were made using Sequencher 4.5 (Gene code corporation) and the sequences were aligned by means of Clustal W .

3.2.3 Expression analysis

Expression of *pros28.1A* was studied in males and females as well as male tissues. Tissues were homogenized in a glass homogenizer and total RNA was extracted as described by the Quiagen protocol from ~30 males and virgin females. Mature males (1-5 days old) were dissected in saline solution for testis and accessory gland and the carcass (gonadosectomized body). The tissues were preserved in RNA-later solution (Ambion) soaked at 4ºC overnight followed by -80ºC until they were processed. RNA was extracted from 20 gonadectomized males, 100 testes and 100 accessory glands of *D. simulans, D. yakuba,* and *D. erecta*.

RT-PCR was conducted on total RNA from males, virgin females, gonadectomized males, testes, and accessory glands. Analysis of expression of intronless genes (such as *pros28.1A*) is challenging

because genomic contamination can produce a band of the same size as that of expected from the cDNA. Therefore, we digested possible contaminating DNA from the total RNA (DNase I amplification grade; invitrogen) and ran controls including DNA digested total RNA without retrotranscriptase. Single strand complementary DNA (cDNA) was synthesized using Superscript (Invitrogen) and oligo-dT (Promega). RT-PCR was carried out using specific primers 5'-GTTCGTGGAGGCAATTGTGTG-3' and 5'-GTACGCCCAGGAAGCTGTTC-3' for amplification of *pros28.1A* in *D. yakuba* and *D. erecta* and 5'-TGCCTGCTAACTAACCCAAAG-3' and 5'-AACTGGGTTAACCTCGAGAAGG-3' in *D. simulans*. *Gapdh2* gene was used as positive control of the RT reaction using 5'-CAAACGAACATGGGAGCATC-3' and 5'-TCAGCCATCAGAGTCGATTC-5' primers.

The full-length 5' end sequence of the *D. melanogaster, D. simulans* and *D. yakuba pros28.1A* transcript was obtained by 5' rapid amplification of cDNA end (RACE) experiments. Total mRNA was treated for 5' phosphate of degraded mRNA, rRNA, tRNA, and genomic DNA leaving the capped mRNA intact. The capped mRNA were then treated to remove the cap and ligated to an adaptor. Single strand cDNA was synthesized from mRNA using Superscript (RLM RACE Ambion) using random decamers as primer. PCRs were conducted to amplify the target transcript using 5'-GCGAGCACAGAATTAATACGACT-3' (RLM RACE Ambion) primer specific to the adapter and primers specific to the *pros28.1A:* 5'-AGGGTCACCTGGTTTTCGAAG-3' , 5'-GGTCACCGGTTTGTCGAAG-3', 5'-GACCTGCCCTCGATTTATTAGGATC-3' was used as 3' primers of *D. melanogaster*, *D. simulans* and *D. yakuba* respectively. Since one round of PCR usually does not yield any product, these products were further amplified using nested primers 5'-CAGCACCACACAATTGGCTCCA-3' specific to *D. melanogaster* and *D. simulans* , and 5'-GTGATCTTGCGCACCGTTCGGTA-3' specific to *D. yakuba* along with RACE Inner primer 5'-CGCGGATCCGAATTAATACGACTCACTATAGG-3' (RLM RACE Ambion). The products were then sequenced. While clear sequencing peaks were observed in *D. malanogaster* and *D. simulans*,suggesting a single TSS, in *D. yakuba*, clear peaks were not observed at the end of the sequence. This is due to multiple TSSs few base pairs apart. Therefore, in the case of he 5'RACE product of *pros28.1A* of *D. yakuba,* the PCR product was cloned using PCR2.1TOPO vector and 10 colonies with insert were sequenced to characterize the ends.

3.2.4 Strains and clones for transformation

Genomic DNA extracted from single *D. melanogaster* fly was used to amplify the putative promoter of *pros28.1A* using 5'CCGCGGATTACTCACCCTAAAC-3', and 5'-AACAATTTGCTTGTGACAAGACCGGT-3'. High fidelity taq (Stratagene) was used to prevent introducing sequence changes in our PCR amplifications. The amplified region includes 5'UTR (56 bp) and 298 bp of the upstream region. This PCR product was TA cloned into PCR2.1TOPO vector (Invitrogen). Digested with SacII (New England Biolabs) and AgeI (New England Biolabs), ligated directionaly to EGFP vector (U55761; CLONTECH CLONE) with T4ligase (New England Biolabs). This cloning places the coding region of EGFP at nearly (17bp) the same position where the coding region of *pros28.1A* is in the genome. The ligated plasmid was then transformed into competent cells (Novagene). After PCR screening and sequencing the plasmid to confirm the integrity of the insert, the purified plasmid using mini prep kit (Quiagen) was digested by AflII (New England Biolabs), blunt ended by Mung Bean (New England Biolabs), and digested by SacII (New England Biolabs). This insert was then ligated to pCaSpeR4 vector (X81645) with a blunt end and a SacII site. The transformed and sequenced plasmid was injected into 30-minute-old embryos along with Turbo transposase to produce insertion of the construct in the genome. Transformants for this construct were obtained after crossing the adults as described below.

The genomic DNA of the above transformant flies was used to PCR the construct with a shorter and shorter promoter region. PCR amplifications were done using 5'CTGCAGTTCGGCTTTCCGTAATTC3', 5'CTGCAGAGTATAATGGCCACGATC3', 5'CTGCAGAATCATTACACTATGGTGTAG3', 5'TTGACTTCAAACTTCAAATGTAAG3', 5'AAATAACTTGTCACAAGCAAATTG3', 5'TTGACTTCAAACTTCAAATGTAACTGGTTAGTG3', 5'GAAAAATTCATATTGTTTCAAGTAAATAACTTG3', primers in the promoter region and using a 3' primer in the P element including EGFP gene and termination signal (5'CTGCAGTGATGAGTTTGGACAAAC3'). High fidelity taq (FINNZYMES) was again used to prevent introducing sequence changes in our PCR amplifications. The amplified sequences were then cloned into PCR2.1TOPO vector (Invitrogen). Digested with KpnI and XhoI it was ligated directionally into pCaSpeR4

vector (X81645) with same sites. These were transformed into XL1Blue competent cells (stratagene). The plasmids were purified and were injected into w118 embryos by Genetics Services.

Flies were maintained in 25℃ in standard corn media. The offsprings were fixed using w;sco/cyo,s and w;2.3/tmbb;sb balancers and w118 for 2nd ,3rd and X chromosomes respectively. The fixed males were dissected in saline solution and images were taken using florescent microscope Olympus BX51TRF setting the UV exposure time manually at one twenty fifth of a second. Controlling the time has allowed us to fine tune for any auto-florescence and enabled us to compare across preparations and lines.

3.2.5 tTAF Strains and crosses

As described in the introduction, tTAFs are testis specific transcription factors that are deployed during spermatogenesis to express genes specifically during this time. We used mutant strains for two of these tTAFs (nht and rye) and try to reveal if *pros28.1A* is directly regulated by these genes or downstream. We cross these mutant lines with our transformed lines aiming to reveal if lack of fluorescence or a decrease in level of fluorescence in the testis of males descending from these crosses could help us in understanding this regulation.

3.4 Results

3.4.1 Pros28.1A expression in different species

Expression of *pros28.1A* was previously shown to be male specific in pupae and adult of *D.*

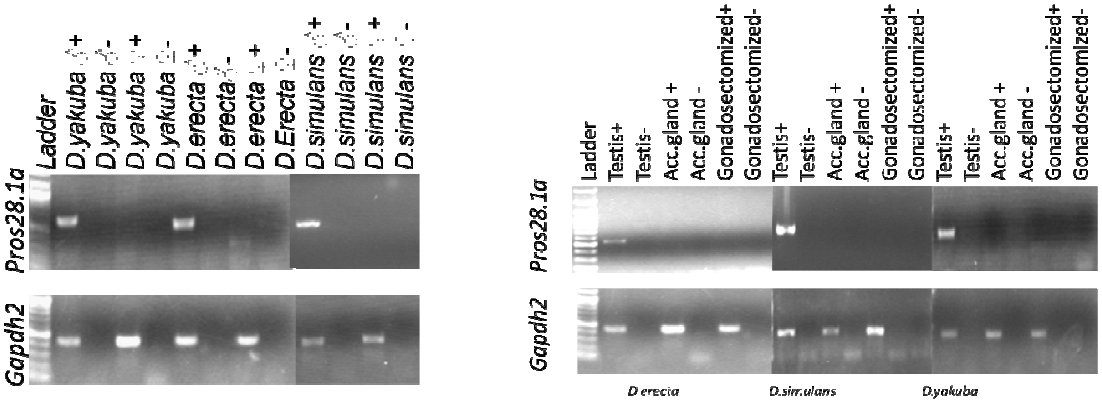

Figure 3.1 *pros28.1A* transcript is testis specific. While the ubiquitously expressed *Gapdh2* is transcribed in female, male and the examined tissues, *pros28.1A* is shown to be male and testis specific.

melanogaster although the parental copy was expressed ubiquitously (Yuan, Miller et al. 1996). Our RT-PCR results also point to the presence of *pros28.1A* transcript only in males of *D. simulans, D. yakuba* and *D. erecta*. The presence of the transcript in males was limited to testis; no transcripts could be detected in the accessory glands and the gonadosectomized body (Figure 3.1).

3.4.2. Transcription Start Site (TSS) recognition

Characterizing regulatory region of a gene (i.e. *pros28.1A*) requires knowledge of the TSS since usually regulatory motifs such as TATA (Ohler, Liao et al. 2002) and β2tubulin promoter (a male specific promoter) (Michiels, Gasch et al. 1989) are found upstream of the TSS.Therefore we performed 5'RACE in *D. melanogaster*, *D. simulans* and *D. yakuba*, the results of which are shown in Figure3.2 in the alignment of *pros28.1A* 5' regions of melanogaster subgroup species (including *D. mauritiana*, *D. santomea, D. teisseri* that were sequenced in the lab). 5'RACE from *D. melanogaster* shows the TSS to be located 56bp upstream of the CDS (i.e. 5'UTR is 56 bp). However this is different in *D. simulans* where the TSS is located 135bp upstream of the predicted CDS, that is 81bp further upstream than in D. melanogaster. In *D. yakuba,* we detected 5 different TSS separated by few base pairs. These different TSSs in different species might reflect the recruitment of different regulatory sequence in these lineages or a shared motif that could be located upstream or downstream depending on the species.

3.4.3 Regulatory region narrow down

Screening the expression pattern of EGFP in the transgenic flies carrying 298bp upstream of the TSS of *pros28.1A* with EGFP as reporter (construct number 1) revealed that the putative promoter driving male specific expression of *pros28.1A* is located in this region (Betrán and Río unpublished). In the effort to narrow down the promoter, five additional constructs were made that contained shorter regions (i.e. 248bp [construct 2], 191bp [construct 3], 83bp [construct 4], 46bp [construct 5] and a construct that lack any upstream sequence from the TSS [construct 6]). This constructs have EGFP as a reporter gene and are described in figure 3.3. As shown in this figure, all the constructs except construct 6 show EGFP expressed in testis of transformant flies. Lack of expression in the 6th construct not only revealed that the putative promoter was located in the 46bp upstream that this construct lacked but also ruled out any possibility that the observed expression pattern was due to leaky expression of EGFP. Also due to the

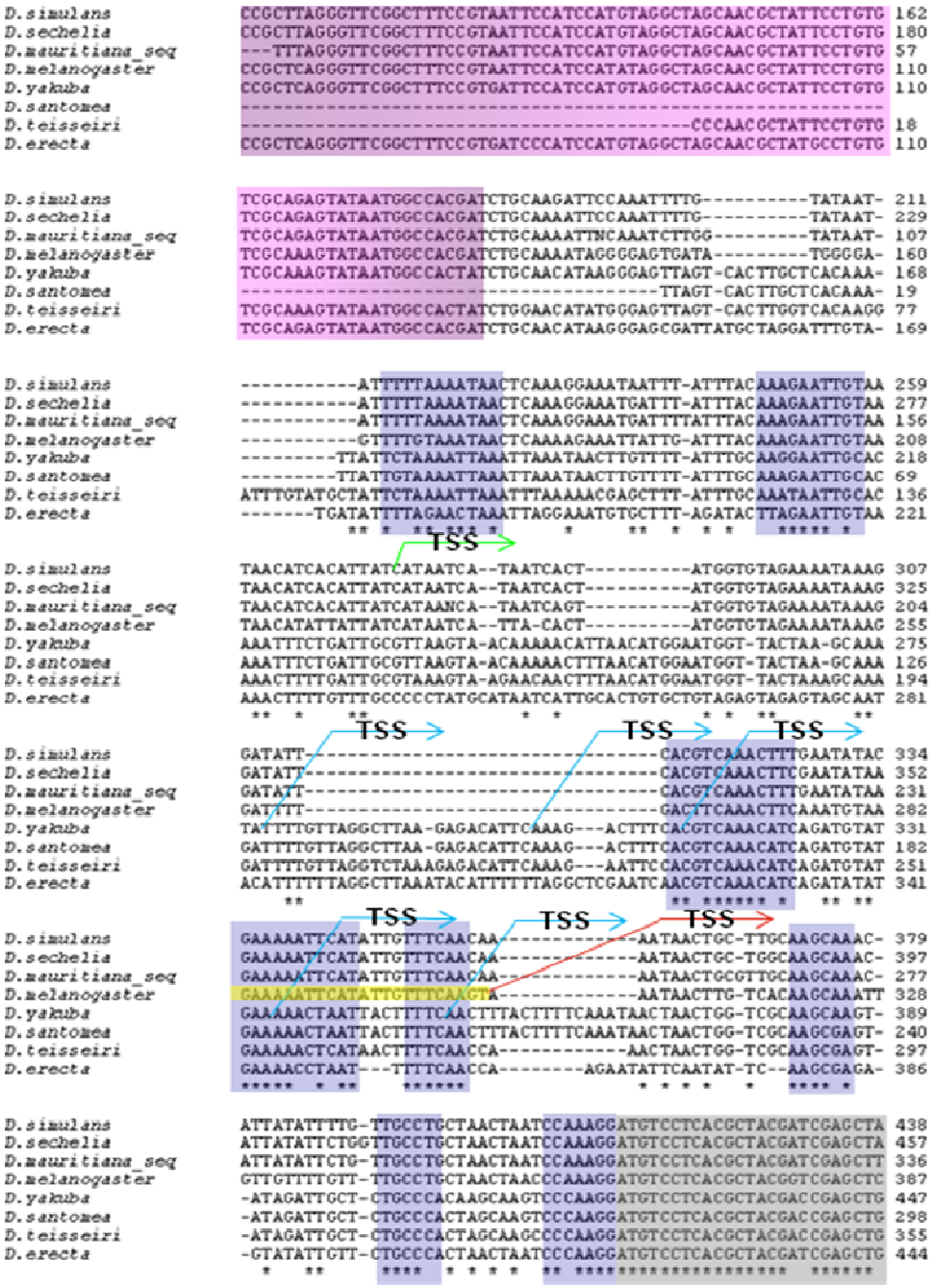

Figure 3.2 Alignment of 5' putative promoter region of *pros28.1A* in melanogaster subgroup. Light grey is the beginning of the coding sequence of *pros28.1A* inserted in the 3rd intron of CG31203, pink is its 4th exon (i.e. CG31203 is encoded in a different strand than *pros28.1A*). Blue boxes are the conserved nucleotides that might have possible roles as regulatory sequence although speculative with different TSS seen in close species(red arrow shows TSS in *D. melanogaster*, green in *D. simulans*, and blue in *D. yakuba*. Yellow box is showing the 23 bp regulatory sequence.

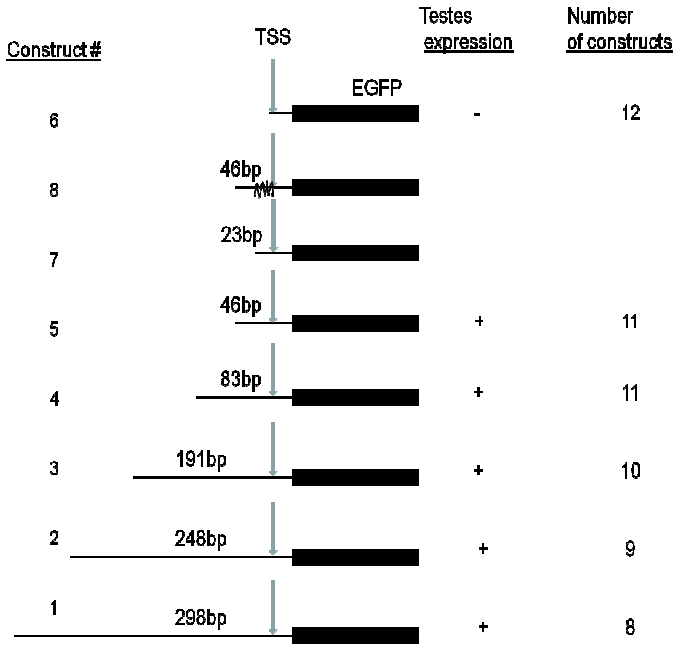

Figure 3.3 An overview of promoter narrow down results

random nature of insertion of these constructs into the genome of *D. melanogaster* minimum 8 independent insertions were screened to confirm the expression of EGFP (5 of which are shown in figure 3.5). As shown in this figure, although the expression intensity varies between independent insertions of the same construct, the pattern of expression is conserved throughout.

Expression level is always compared to w118 adult males auto-fluoresce as negative control. In order to differentiate this florescence with the one derived from expression of EGFP, the basal exposure time was manually fixed to a level shown in figure 3.3 (as described in materials and methods). Comparing this expression with the one with the construct with no upstream sequence the expression was visually equal.

Figure 3.4 Testis auto-fluorescence level in control (w118)

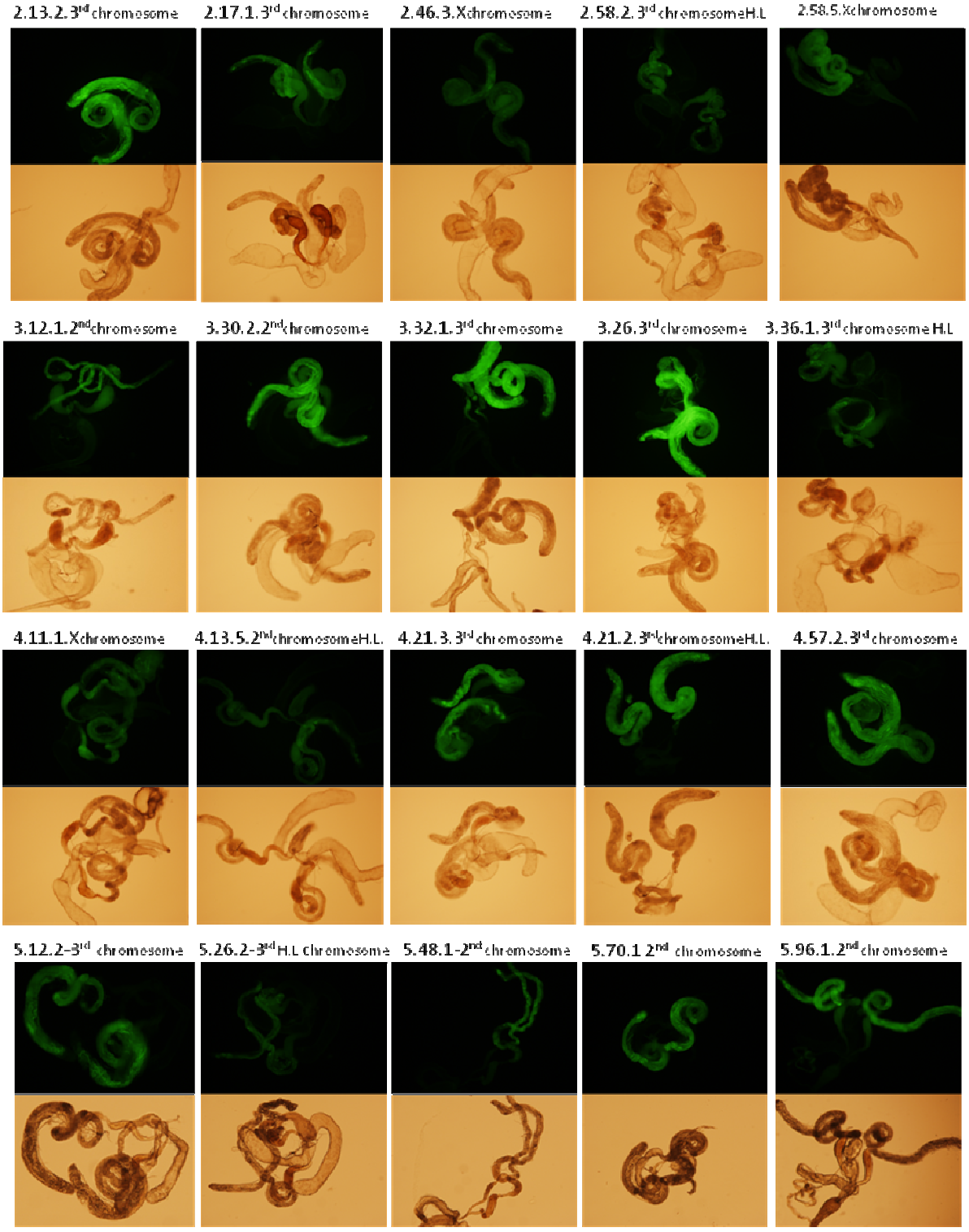

Figure 3.5 Expression of EGFP in constructs 2-5. under exposure of UV light (on top). The corresponding image with white light is shown below it. The numbers shown are the construct number, the independent insertion, the individual and the chromosome where the insert was mapped.

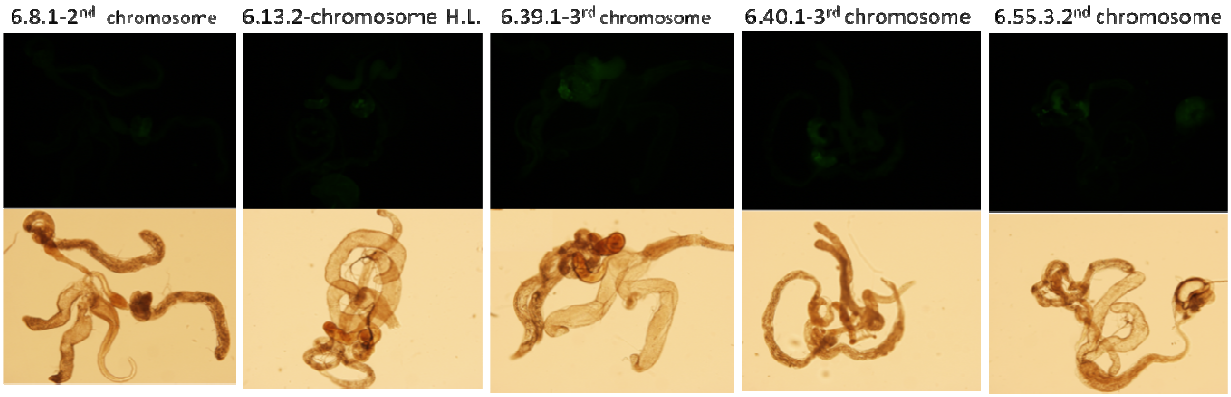

Figure 3.6 Lack of EGFP expression in constructs 6. Construct 6 lacks the expression of the EGFP with the level of florescence comparable to the auto-fluorescence level in w118 males (control).

The alignment of the 46bp sequence driving expression of EGFP in testis of transgenic flies in species of melanogaster subgroups shows three conserved regions. In order to identify which region or combination of regions might be responsible for the pattern of expression observed, two additional constructs were made: construct 7 (Region I-absent) lacking the first 23bps where the Region I resides in and the other construct (construct 8; Region II-III-mutated) was made mutating the Region II and the Region III to random sequences as shown in figure3.7. As shown in figure3.8, the sequence driving the expression of the EGFP in transgenic flies is the 23bp motif 5'-GAAAATTCATATTGTTTCAAGT-3'.

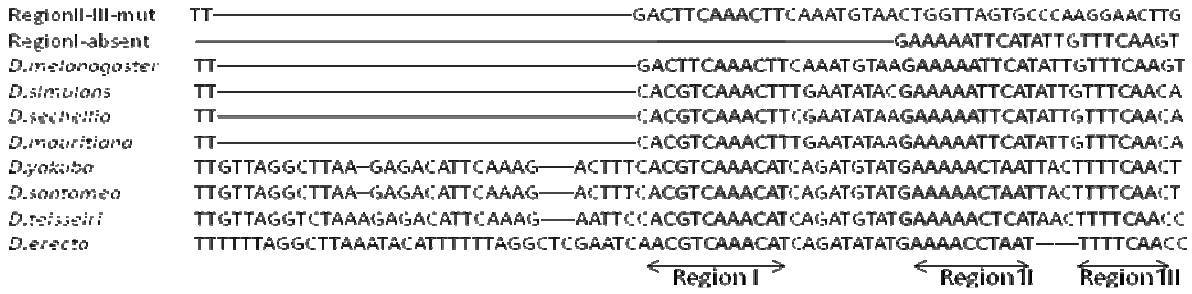

Figure 3. 7- Alignment of the 46bp upstream of TSS in *pros28.1A* driving expression of EGFP in testis of transformant flies. Grey boxes show the conserved regions and the two sequences on top show the sequences present in constructs 7 and 8 used to further explore the regulatory element

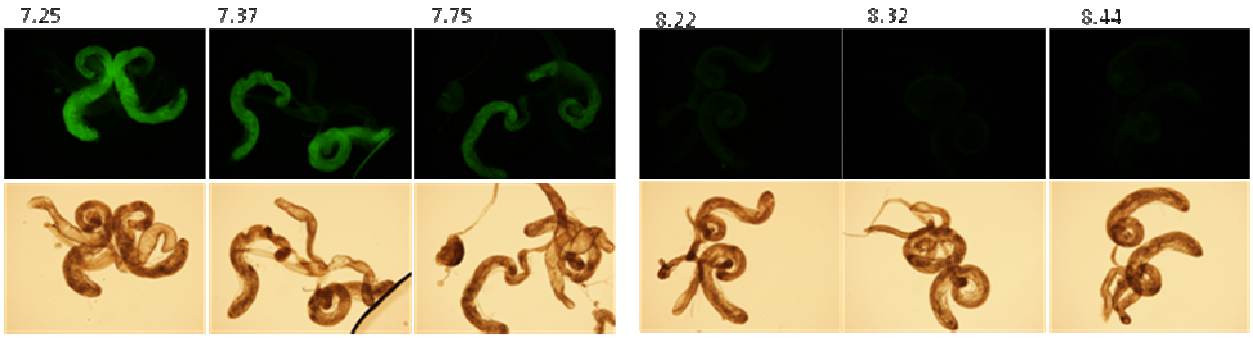

Figure 3.8 Expression of EGFP in constructs 7 and 8. Expression of EGFP in construct 7 and absence of this in the 8th construct points to the 23 bp responsible for driving the expression of transgene and thus the *pros28.1A.*

3.4.4 Testis specific promoter

To check if this 46 bp region carries a testis specific promoter, expression of the construct with the 46 bp region was checked in different tissues in larvae and adults in several lines. Since insertions are random, some might reside in euchromatic regions where higher expression of EGFP could be seen or in regions such as heterochromatic where the expression is less obvious. This is also the case in screening the expression of EGFP in testis of individuals with insertions in the X chromosome or the ones that are homozygous lethal. To be able to distinguish the pattern of expression better, the 3 insertions which showed the highest expression level in testis were chosen (5.12.2, 5.70.2, and 5,96.2). The expression was screened in the 3rd instar larvae in gonads and gut and in adults in abdominal tissues such as gonads (ovary, testis, accessory glands) and gut and was compared to the w118 strain. As shown in Figure3.9 the tissues do not show fluorescence apart from some auto-fluorescence that is detected in the control as well with an exception of the male gonads in larvae and adults. This reveals that the 46 bp drive testis specific expression as expected from our RT-PCR results.

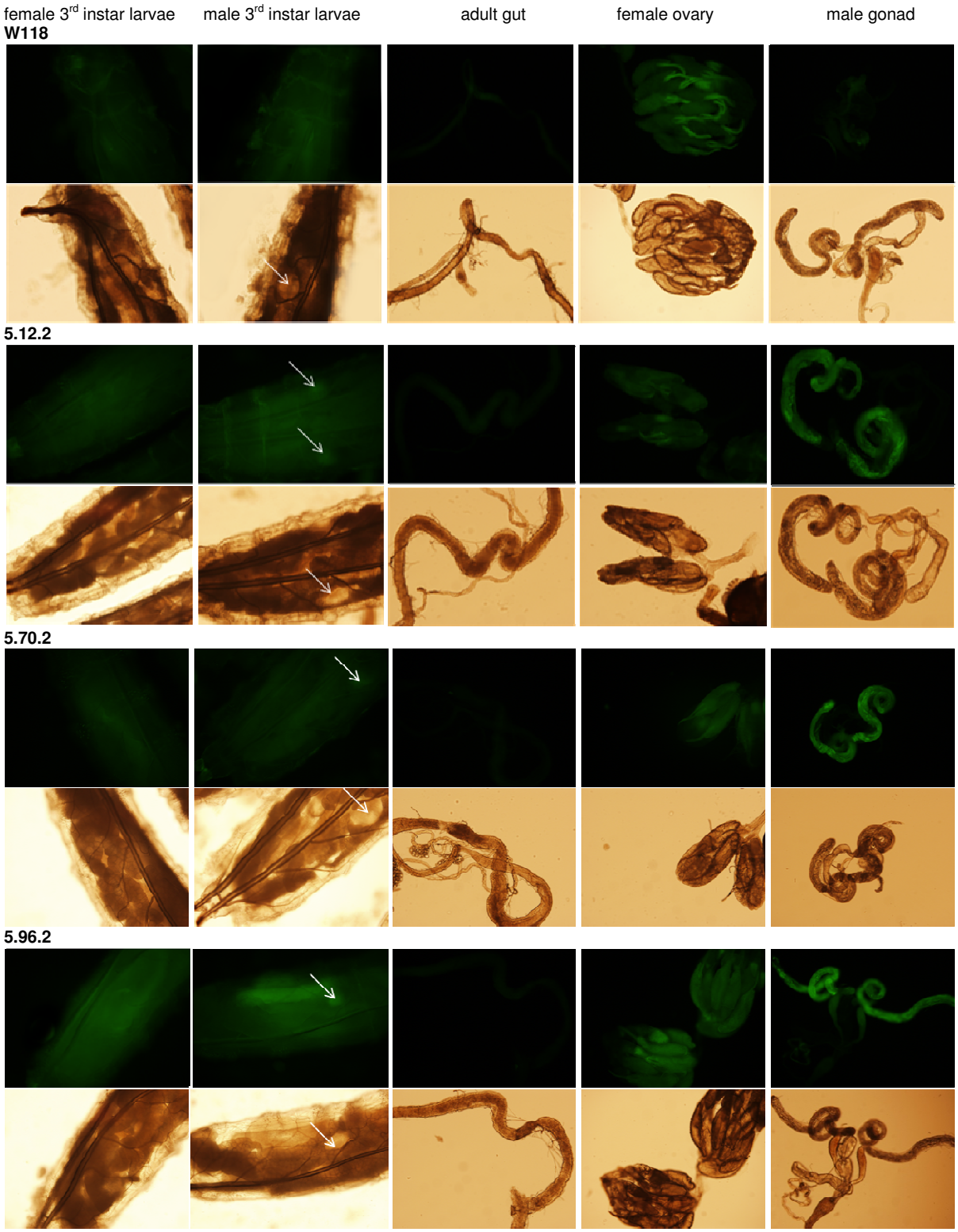

Figure 3.9. Expression of EGFP in the construct with 46bp putative promoter points to a testis specific nature of the promoter present. Arrows show the male gonad. Female gonad in larvae is not detectable so in cases of not observing the gonads, the individual was counted as female

3.4.5 Putative regulatory region: its quality and origin

Manual checking for the presence of known motifs in this 46bp reveal no similarities. We looked for promoter elements described by (Ohler, Liao et al. 2002; FitzGerald, Sturgill et al. 2006), the PACE element regulating expression of proteasomic genes in yeast (Mannhaupt, Schnall et al. 1999) and the *β2-tubulin* testis specific promoter (Michiels, Gasch et al. 1989) and failed to detect a known or similar to a known motif responsible for transcription of *pros28.1A* in *D. melanogaster*.

In studying the origin of this regulatory region and whether it was recruited from the intron of CG31203, parental gene or any sequence inserted at the time of insertion following studies were performed. Alignment of *CG31203* 3rd intron in the species lacking *pros28.1A* demonstrates a short and poorly conserved intron (figure3.10) revealing the unlikely recruitment of the regulatory sequence from this intron.

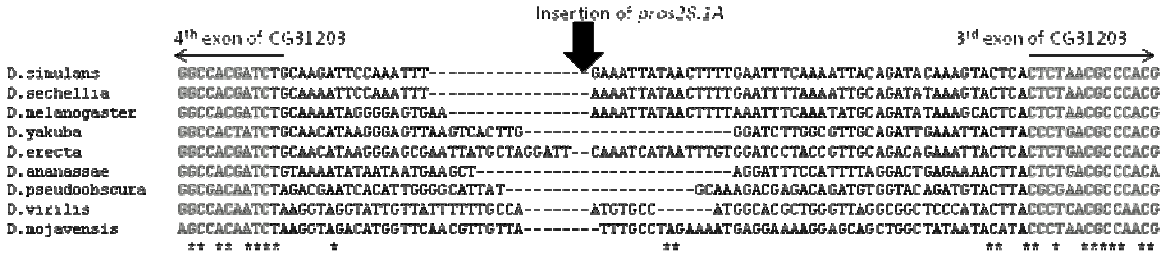

Figure 3.10- Sequence alignment of *CG31203* intron in several Drosophila species. Grey is the exons of CG31203 and arrow points to the approximate position where the pros28.1A and likely additional sequences at the 5' end were inserted.

There is ~157bp region located upstream of the *pros28.1A* TSS that is not present in the species lacking this insertion. Alignment of the 23 bp regulatory region with parental upstream region reveals little similarity as shown in figure 3.11 (11 substitutions and one deletion out of 23bp). This is when the same region is extremely conserved when compared to the corresponding sequence in *D. simulans* (one substitution out of 23bps). Thus this region was not originated from the parental sequence but could have diverged from parts of transcript carried over.

Also the fact that the parental UTR is 146 bp whereas the sequence upstream of the retrogene's coding sequence to the intron of CG31203 is 157bp brings up the possibility that this region originated at the time of insertion as a transduced sequence. There are reports suggesting the ability of the LINE

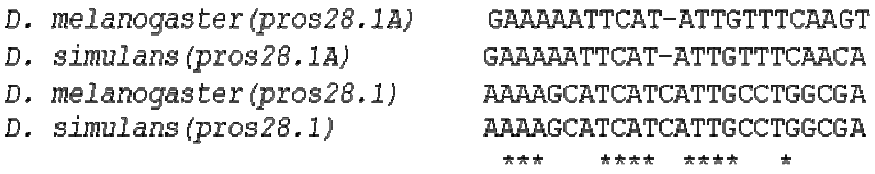

Figure 3.11 23 bp regulatory sequence against close species and parental gene. Sequence alignment of the 23bp motif between *D. melanogaster* and *D. simulans* compared with the corresponding region in parental gene in the same species.

element to jump from one template to another thus introducing a piece of DNA that was not present in the original transcript examples of this are integration of extra pieces of LINE element or the flanking sequences (Babushok, Ostertag et al. 2006). In order to reveal the origin of if this region, we blasted this sequence and found sporadic insignificant hits against the sequenced genomes. From these findings, we concluded that *pros28.1A* recruited a *"de novo"* regulatory sequence.

3.4.6. tTAFs involvement in the regulation of pros28.1A

Upon recognizing the 46bps that harbor a testis specific promoter motif, we made crosses between testis specific Transcription Associated Factor (tTAF) mutant straints (rye1, rye2 and nht) and our EGFP transgene to look at the changes in fluorescence level due to absence of transcription. In theory if this testis specific transcription machinery using tTAFs is responsible for transcribing the *pros28.1A* or is upstream in the cascade, the absence of one allele of the tTAF might lower or abolish expression of the EGFP. Crossing the tTAF mutants (nht, rye1 ,and rye2) with the transformant flies, the offspring will have one copy of the construct in one chromatid against the tTAF mutant on the other chromatid. As a result the maximum expression of the EGFP would be half the level of the fixed individuals. Thus the comparison for the level of expression was made with heterozygotes for the EGFP transgene. Figure 3.12 shows the results. We still observe EGFP expression in testis of these individuals. That might mean that tTAF are not involved or that one copy is enough to show the same levels of EGFP. Quantification of fluorescence under a confocal microscope and/or crosses where tTAF mutatnts are in homozygotes are needed to reveal any involvement of tTAFs in transcription of *pros28.1A.*

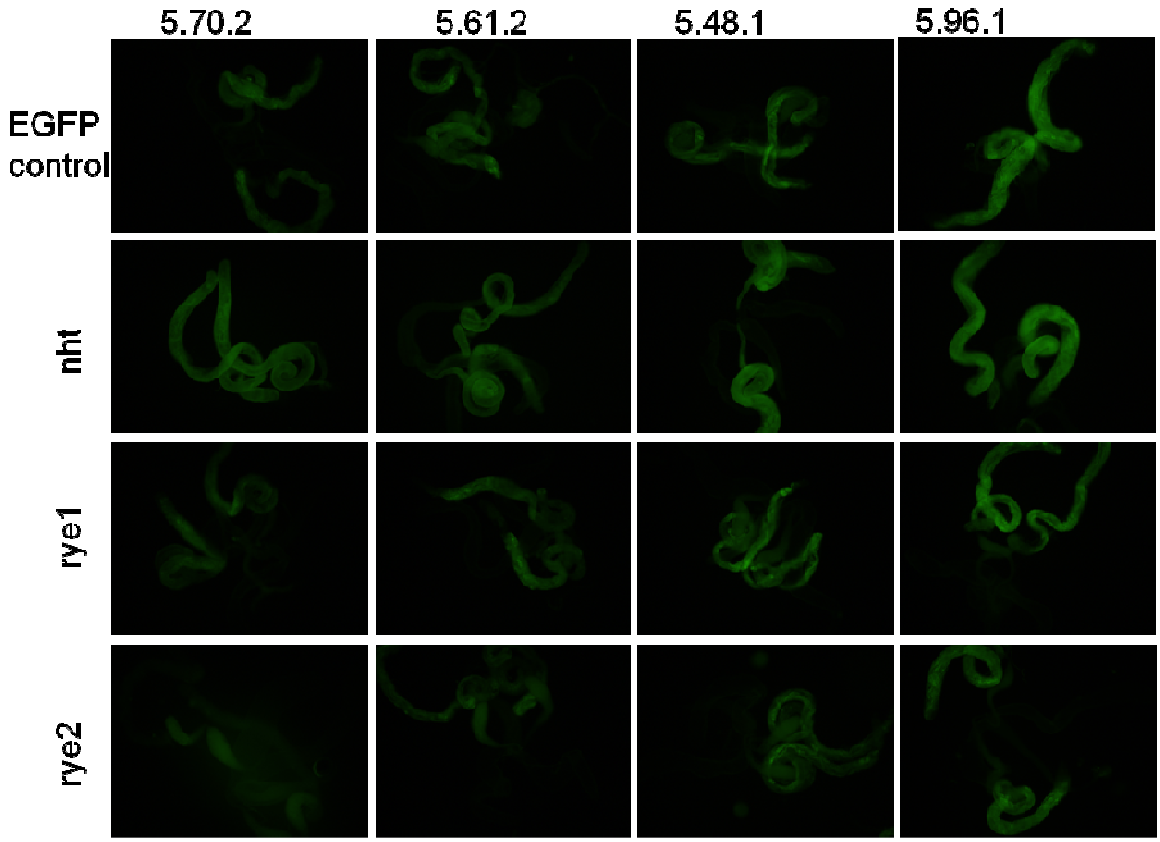

Figure 3.12 Florecence in testis of the tTAF and EGFP transgene progeny. First row shows the expression level of the constructs with insertion in hetrerozygosity.

3.4 Discussion

Promoters are shown to contain information that directs tissue specific mRNA expression (Bielinska, Lu et al. 2005). For instance presence of TATA motif in a promoter positively correlates with gene expression in somatic tissue but negatively correlates with expression in germline tissue (FitzGerald, Sturgill et al. 2006). However, promoter of many of the testis expressed genes has shown to be short and simple sequences in close vicinity of TSS (Michiels, Gasch et al. 1989; Zhang, Sturgill et al. 2007). Our data supports this record: we have shown that 23bp upstream of the TSS is enough to drive male specific expression of *pros28.1A* in *D. melanogaster*. Interestingly but puzzlingly, *D. simulans* transcript of *pros28.1A* begins upstream of the one in *D. melanogaster* and it is therefore likely to use a different regulatory region. This likely reflects the high turnover and frequent testis regulatory region recruitment and would be in agreement with the observation that testis expression is the most rapidly

changing feature of the Drosophila transcriptome (Zhang, Sturgill et al. 2007). We are now making constructs that include the upstream region of *D. simulans* to reveal if there is another testis specific motif there.

Characterizing a male specific promoter not only helps in better understanding the recruitment of such promoters from the sequence available, but also helps in recognizing the transcription machinery involved. Due to the expression overlap of testis specific TAFs and the *pros28.1A*, involvement of the tTAFs in transcription of *pros28.1A* is expected. More work needs to be done to investigate this possibility although our preliminary results shows no evidence of tTAFs involvement. Narrowing down to the motif, it is also interesting to know if the same motif, or its degenerate forms are present in other testis specific genes. Simple blast analyses of the 46 bps does not detect this, however use of other softwares designed to find motifs (i.e. Patser;(Hertz and Stormo 1999) should also be used in the future.

CHAPTER 4

FUNCTIONAL ANALYSES OF PROTEASOME PROTEIN RETROGENES IN D. MELANOGASTER

4.1 Summary

In this chapter, I present preliminary analyses that seek to better understand the role and function of two of the proteasome duplicates using RNA interference (RNAi). We tried to understand the effects of absence of *prosα4T1,and prosα4T2* transcript during spermatogenesis and use *prosα6T* as positive control.

4.2 Retrogenes analyzed and functional approaches

The ability to express or suppress a gene in a directed fashion is a useful tool to analyze its role. One way to detect the phenotype caused by the absence of the transcript is the RNAi mediated decay of a targeted transcript. The double stranded RNA is recognized and cleaved into 21-26 nucleotide RNAs by the Dicer complex . These small interfering RNAs (siRNAs) will then target complementary mRNAs for degradation by RISC which is an RNA-induced silencing complex. In a genome wide transgenic RNAi library generated in 2007 by Dietzl, Chen et al., 88% of the predicted protein-coding genes in the *Drosophila* genome have at least one RNAi line capable of silencing its transcript These RNAi transgenes are short inverted repeat gene fragments transformed into *Drosophila* under the control of the UAS promoter (Dietzl, Chen et al. 2007). The protein coded by GAL4 binds to the UAS promoter and activates expression(Brand and Perrimon 1993) of transgene. This allows for studying the effects of absence of a transcript in a tissue or stage specific manner by manipulating the expression of GAL4. . .

RNAi lines for *prosα4T1*, *prosα4T2* and *prosα6T* were ordered from the Vienna *Drosophila* Stock Center (VDRC) (Dietzl, Chen et al. 2007). Four GAL4 driving lines (T80, T100, C564 and C855) that drive the RNAi constructs at different stages in spermatogenesis (Hrdlicka, Gibson et al. 2002) were kindly provided by Norbert Perrimon along with a UAS-mcD8EGFP line. Although all the driver lines were reported to show testis expression of GAL4 in different stages of spermatogenesis, these lines were

further screened in the lab. Crossing the 4 lines with the UAS-mcD8GFP was performed such that the progeny have a copy of GAL4 driver and UAS-mCD8GFP. As shown in figure 4.1, T80 drives expression of GFP in the testis sheath while the others drive expression of GFP after meiosis and during the later stages of spermatogenesis (i.e individualization and elongation). To better observe these expression patterns all images but T80 were taken with longer exposure time (1/3 of a second). The genes under study begin to express earlier than this stage as well as during individualization and elongation (see Chapter 3). Therefore, we have essayed them to see if driving the RNAi constructs has male fertility effects.

Females homozygous for GAL4 driver line was crossed with males homozygous for RNAi construct for *prosα4T1, prosα4T2* and *prosα6T*. *Prosα6T* is regarded as a positive control because the mutants of *prosα6T* were shown to be male sterile (Zhong and Belote 2007). These flies were then incubated at 29ºC to enhance the GAL4 activity. The progeny of this cross marked by red color of the eye were then screened for fertility by being crossed with the virgin individuals from the original RNAi stock in 5 replicates.

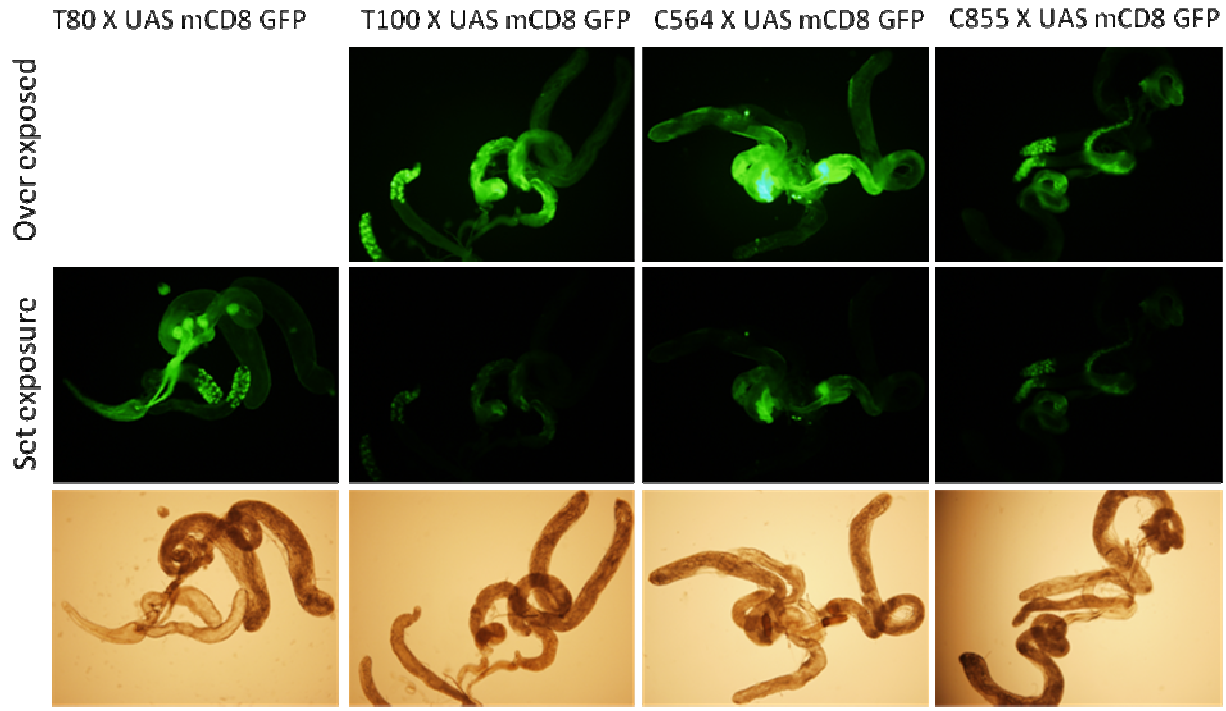

Figure 4.1 Expression of EGFP in testis of individuals with a copy of GAL4 transgene and UASmCD8 GFP. To better view the localization of GFP, the images were taken with 8X longer exposure time (shown on top).

4.3 Results and discussion

As explained earlier, the GAL4 drivers were expressing little later than expected thus all the crosses including the controls were shown to be fertile. In future GAL4 drivers will be obtained to either drive the expression of transgene ubiquitously or in a testis specific manner (i.e. using the 23bp promoter of *prosα4t*).

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

Duplication of proteasomal genes has been shown to be common phenomena in plants, mammals and *Drosophila*. Previously over one third of the genes encoding for proteasome in *Drosophila* was recognized to have duplicate forms with male specific expression. Due to male specific expression pattern of all the known isoforms in *D. melanogaster*, it is very likely that there is a testis specific proteasome in the testis substituting the somatic one.

In this work, we have shown that almost all genes encoding for the core particle (20S subunit) had given rise to at least one duplicate. This is when the lid (19S) has very few duplicates. This difference might be due to the fact that the core particle is the main catatalytic domain, and it might be advantageous to have duplicate forms during the sperm individualization and maturation when an extensive remodeling and condensing occurs to cleave the proteins in a different fashion. In the study where the cellular localization of *prosα6T* and *prosα3T* duplicates was studied, these transcripts were more localized to nucleus and in the head of the sperm. In *Drosophila*, the transcription timing of these duplicates overlaps with histone-protamine and protamine- histone transition. This might suggest a role of the testis specific proteasome in chromatin remodeling during condensation and probably after fertilization (Zhong and Belote 2007).

While the 20S core particle with more duplicates is possibly a more heterogeneous complex, the 19S regulatory cap is shown to be more structurally conserved. As shown previously, only few subunits that interact with the core particle have male specific duplicates. This could be due to preserved function of the remaining subunits such as the ones that are responsible for binding to the ubiquitin, unfolding the protein and feeding the unfolded protein to the core particle.

Looking at the evolution of some of these duplicates we showed that while all of the studied retroposed copies evolve faster than the parental genes with some evolving extremely fast and others

moderately fast, more data is needed to reveal how often positive selection has been acting, in what residues and in what lineages.

Finally, we showed that a short motif quite close to the TSS is enough to drive the expression of α4 retroposed copy (*prosα4T1*) in *D. melanogaster.* While this *"de novo"* promoter showed to be one of the few male specific promoters, it might not be used in other close related species (i.e *D. simulans* and *D. yakuba*). If this will be the case the evolution of the testis specific promoters are going to be much faster than previously thought.

5.2 Future works

This research along with other researchers results lead us to many additional questions:

1. How do other duplicates evolve when compared to their parental gene?
2. Is there any positive selection on different sites of these duplicates? Could this be detected using the PAML branch site models between the species or other tests like the McDonald-Kreitman is needed in different populations of the same species to answer this question?
3. Can we get more evidence of a specialized proteasome and of neofunctionalization of some paralogs?
4. How often do we find short promoters in testis-expressed retrogenes? What is their quality and origin?
5. Are short promoters a general feature for testis expression?
6. What transcription machinery is responsible for the recognition of the described motif? Is this machinery also specific to testis of Drosophila?
7. Since the whole sperm enters the egg in Drosophila, what happens to these proteasome proteins after fertilization?

REFERENCES

(2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." **447**(7146): 799-816.

Anne Davy, P. B., Nicolas Thierry-Mieg, Philippe vaglio, Joseph Hitti, Lynn Doucette-Stamm, Danielle Thierry-Mieg, Jerome Reboul, simon Boulton, Albertha J. M. Walhout, Olivier coux, Marc Vidal (2001). "A protein-protein interaction map of the Caenorhabditis elegans 26S proteasome." Embo J **2**(9): 821-828.

Ashburner, M., K. Golic, et al. (2004). Drosophila. cold Spring Harbor, Newyork, Cold spring Harbor Labratory press.

Babushok, D. V., E. M. Ostertag, et al. (2006). "L1 integration in a transgenic mouse model." Genome Res. **16**(2): 240-250.

Bai, Y., C. Casola, et al. (2007). "Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila." Genome Biol **8**(1): R11.

Belote, J. M., M. Miller, et al. (1998). "Evolutionary conservation of a testes-specific proteasome subunit gene in Drosophila." Gene **215**(1): 93-100.

Belote, J. M. and L. Zhong (2005). "Proteasome gene duplications in mammals, flies and plants." Genes and genomes **1**: 107-129.

Betran, E., J. J. Emerson, et al. (2004). "Sex chromosomes and male functions: where do new genes go?" Cell Cycle **3**(7): 873-5.

Betran, E. and M. Long (2002). "Expansion of genome coding regions by acquisition of new genes." Genetica **115**(1): 65-80.

Betran, E. and M. Long (2003). "Dntf-2r, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection." Genetics **164**(3): 977-88.

Betran, E., K. Thornton, et al. (2002). "Retroposed new genes out of the X in Drosophila." Genome Res **12**(12): 1854-9.

Bielinska, B., J. Lu, et al. (2005). "Core Promoter Sequences Contribute to ovo-B Regulation in the Drosophila melanogaster Germline." Genetics **169**(1): 161-172.

Brand, A. and N. Perrimon (1993). "Targeted gene expression as a means of altering cell fates and generating dominant phenotypes." Development **118**(2): 401-415.

Brosius, J. (1999). "RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements." Gene **238**(1): 115-134.

Chen, X., M. Hiller, et al. (2005). "Tissue-Specific TAFs Counteract Polycomb to Turn on Terminal Differentiation." Science **310**(5749): 869-872.

Ciechanover, A. and A. L. Schwartz (1998). "The ubiquitin-proteasome pathway: the complexity and myriad functions of proteins death." Proc Natl Acad Sci U S A **95**(6): 2727-30.

Clark, A. G., M. B. Eisen, et al. (2007). "Evolution of genes and genomes on the Drosophila phylogeny." Nature **450**(7167): 203-18.

Coux, O. (2003). "An interaction map of proteasome subunits." Biochem. Soc. Trans. **31**(2): 465-469.

Dietzl, G., D. Chen, et al. (2007). "A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila." **448**(7150): 151-156.

Emerson, J. J., H. Kaessmann, et al. (2004). "Extensive gene traffic on the mammalian X chromosome." Science **303**(5657): 537-40.

Emerson, J. J., H. Kaessmann, et al. (2004). "Extensive Gene Traffic on the Mammalian X Chromosome." Science **303**(5657): 537-540.

Esnault, C. c., J. l. Maestre, et al. (2000). "Human LINE retrotransposons generate processed pseudogenes." **24**(4): 363-367.

Feschotte, C. (2008). "Transposable elements and the evolution of regulatory networks." **9**(5): 397-405.

Feschotte, C. and E. J. Pritham (2007). "DNA transposons and the evolution of eukaryotic genomes." Annu Rev Genet **41**: 331-68.
FitzGerald, P. C., D. Sturgill, et al. (2006). "Comparative genomics of Drosophila and human core promoters." Genome Biol **7**(7): R53.
Fraser, C. M., J. D. Gocayne, et al. (1995). "The Minimal Gene Complement of Mycoplasma genitalium." Science **270**(5235): 397-404.
Gaczynska, M., K. L. Rock, et al. (1993). "[gamma]-Interferon and expression of MHC genes regulate peptide hydrolysis by proteasomes." **365**(6443): 264-267.
Garcia-Lora, A., I. Algarra, et al. (2003). "MHC class I antigens, immune surveillance, and tumor immune escape." J Cell Physiol **195**(3): 346-55.
Glickman, M. H. and A. Ciechanover (2002). "The Ubiquitin-Proteasome Proteolytic Pathway: Destruction for the Sake of Construction." Physiol. Rev. **82**(2): 373-428.
Graur D, L. W. H. (1999). Fundamentals of Mollecular evolution. Sunderland, Massachusetts, Sinauer Associates Inc.
Grisham, R. G. a. c. (1998). Biochemistry, Harcourt Inc.
Groettrup, M., M. van den Broek, et al. (2001). "Structural plasticity of the proteasome and its function in antigen processing." Crit Rev Immunol **21**(4): 339-58.
Groll, M., M. Bajorek, et al. (2000). "A gated channel into the proteasome core particle." **7**(11): 1062-1067.
Groll, M., L. Ditzel, et al. (1997). "Structure of 20S proteasome from yeast at 2.4A resolution." **386**(6624): 463-471.
Groll, M. and R. Huber (2003). "Substrate access and processing by the 20S proteasome core particle." Int J Biochem Cell Biol **35**(5): 606-16.
Hass C, P.-H. B., Multhaup G, Beyreuther K, Kloetzel PM. (1990). "The Drosophila PROS-28.1 gene is a member of the proteasome gene family." Gene **90**(2): 235-41.
Hertz, G. and G. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." Bioinformatics **15**(7): 563-577.
Hiller, M. A., T.-Y. Lin, et al. (2001). "Developmental regulation of transcription by a tissue-specific TAF homolog." Genes Dev. **15**(8): 1021-1030.
Hochheimer, A. and R. Tjian (2003). "Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression." Genes Dev **17**(11): 1309-20.
Hrdlicka, L., M. Gibson, et al. (2002). "Analysis of twenty-four Gal4 lines in Drosophila melanogaster." Genesis **34**(1-2): 51-7.
Kalmykova, A. I., D. I. Nurminsky, et al. (2005). "Regulated chromatin domain comprising cluster of co-expressed genes in Drosophila melanogaster." Nucl. Acids Res. **33**(5): 1435-1444.
Kanhere, A. and M. Bansal (2005). "Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes." Nucl. Acids Res. **33**(10): 3165-3175.
Kazazian, H. H., Jr. (2004). "Mobile Elements: Drivers of Genome Evolution." Science **303**(5664): 1626-1632.
Kutach, A. K. and J. T. Kadonaga (2000). "The Downstream Promoter Element DPE Appears To Be as Widely Used as the TATA Box in Drosophila Core Promoters." Mol. Cell. Biol. **20**(13): 4754-4764.
Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-8.
Levine, M. T., C. D. Jones, et al. (2006). "Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression." Proceedings of the National Academy of Sciences **103**(26): 9935-9939.
Lewin, B. (2004). Gene VIII, Pretice-Hall.
Lynch, M. and A. Force (2000). "The probability of duplicate gene preservation by subfunctionalization." Genetics **154**(1): 459-73.
Ma, J., E. Katz, et al. (2002). "Expression of proteasome subunit isoforms during spermatogenesis in Drosophila melanogaster." Insect Molecular Biology **11**(6): 627-639.

Mannhaupt, G., R. Schnall, et al. (1999). "Rpn4p acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26S proteasomal and other genes in yeast." FEBS Letters **450**(1-2): 27-34.
McCarrey, J. R. (Jan., 1994). " Evolution of Tissue-Specific Gene Expression in Mammals." BioScience **44**(1): 20-27.
McCarrey, J. R., C. B. Geyer, et al. (2005). "Epigenetic regulation of testis-specific gene expression." Ann N Y Acad Sci **1061**: 226-42.
McCarrey, J. R., C. Watson, et al. (2002). "X-chromosome inactivation during spermatogenesis is regulated by an Xist/Tsix-independent mechanism in the mouse." Genesis **34**(4): 257-66.
Michiels, F., A. Gasch, et al. (1989). "A 14 bp promoter element directs the testis specificity of the Drosophila beta 2 tubulin gene." Embo J **8**(5): 1559-65.
Ohler, U., G. C. Liao, et al. (2002). "Computational analysis of core promoters in the Drosophila genome." Genome Biol **3**(12): RESEARCH0087.
Richler, C., H. Soreq, et al. (1992). "X inactivation in mammalian testis is correlated with inactive X-specific transcription." Nat Genet **2**(3): 192-5.
Smale, S. T. and J. T. Kadonaga (2003). "THE RNA POLYMERASE II CORE PROMOTER." Annual Review of Biochemistry **72**(1): 449 LP - 479.
Spofford, J. B. (Jul. - Aug., 1969). " Heterosis and the Evolution of Duplications." The American Naturalist **103**(932): 407-432.
Toba, G. and T. Aigaki (2000). "Disruption of the Microsomal glutathione S-transferase-like gene reduces life span of Drosophila melanogaster." Gene **253**(2): 179-187.
Torgerson, D. G. and R. S. Singh (2004). "Rapid evolution through gene duplication and subfunctionalization of the testes-specific alpha4 proteasome subunits in Drosophila." Genetics **168**(3): 1421-32.
Unno, M., T. Mizushima, et al. (2002). "The Structure of the Mammalian 20S Proteasome at 2.75 Å Resolution." Structure **10**(5): 609-618.
Wilson, R. J., J. L. Goodman, et al. (2008). "FlyBase: integration and improvements to query tools." Nucleic Acids Res **36**(Database issue): D588-93.
Yang, P., H. Fu, et al. (2004). "Purification of the Arabidopsis 26 S Proteasome: BIOCHEMICAL AND MOLECULAR ANALYSES REVEALED THE PRESENCE OF MULTIPLE ISOFORMS." J. Biol. Chem. **279**(8): 6401-6413.
Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Comput Appl Biosci **13**(5): 555-6.
Yuan, X., M. Miller, et al. (1996). "Duplicated proteasome subunit genes in Drosophila melanogaster encoding testes-specific isoforms." Genetics **144**(1): 147-57.
Zhang, Y., D. Sturgill, et al. (2007). "Constraint and turnover in sex-biased gene expression in the genus Drosophila." **450**(7167): 233-237.
Zhong, L. and J. M. Belote (2007). "The testis-specific proteasome subunit Prosalpha6T of D. melanogaster is required for individualization and nuclear maturation during spermatogenesis." Development **134**(19): 3517-25.

BIOGRAPHICAL INFORMATION

Mehran was born in 1981 in Shiraz, Iran. Graduated from high school she migrated to United States October 2001, and took English as Second Language classes. In less than two years, she joined local community college and finished her English and college prerequisites joining University of Texas at Arlington Fall of 2004. Received her B.S. degree in Biology spring 2006 with Honors, she joined the joined the Masters program, next fall. Had begun working in Betrán lab for her undergraduate thesis, she continued her work on recognizing the promoter driving the expression of pros28.1A and expanded her work to other proteasomal duplicates and their mode of evolution. She received her Masters in Biology with focus on Genomics.