Rcd-1 RELATED: A POSITIVELY SELECTED RETROGENE WITH SPERMATOGENESIS FUNCTION IN *DROSOPHILA*

By

TANIYA MULIYIL

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

DECEMBER 2007

ACKNOWLEDGEMENTS

I wish to express my deep sense of gratitude to my supervising professor Dr. Esther Betrán who guided my work with scholarly advice and warm personal affection.

I am deeply indebted to Dr. Shawn Christensen and his lab members for providing the facilities for the successful completion of part of the work.

I also owe sincere thanks to Dr. Claudio Casola, Xavier Río, Mansi Motiwale, Mehran Sorourian, Charles Tracy and Dr. Chitra Chandrashekaran for their help for the completion of this work.

I would like to thank you to all graduate students, faculty and staff of the Department of Biology, for your support and friendship.

I am also thankful to Ratheesh Nair and all my family members for their love, encouragement and help during difficulties.

November 26, 2007

ABSTRACT

Rcd-1 RELATED: A POSITIVELY SELECTED RETROGENE WITH SPERMATOGENESIS FUNCTION IN *DROSOPHILA*

Publication No. ______

TANIYA MULIYIL, M.S. Biology

The University of Texas at Arlington, 2007

Supervising Professor: Dr Esther Betrán

Gene duplication is one of the major forces driving genome evolution. Our study focuses on a particular retrogene Rcd-1 related (Rcd-1r) that originated through retroposition of the parental gene Required for cell differentiation 1 (Rcd-1) about 5-13 million years ago. Rcd-1r is present in D. melanogaster, D. simulans, D. sechellia and D. mauritiana whereas the parental gene is present in all the Drosophila species examined. Rcd-1r is inserted in the 3' UTR of the gene CG13102 and it is encoded in the complementary strand. Previous studies by Castrillion et al. showed that a P element inserted in the 5' UTR of Rcd-1r (that it is also the 3' UTR of CG13102) resulted in male sterility. The aim of this study was to find out which of the two genes CG13102 or

Rcd-1r is responsible for the sterility phenotype associated with the P element insertion and to study Rcd-1r molecular evolution.

We have studied the quality of the transcripts and quantified the expression of CG13102 and Rcd-1r in both wild type and sterile males. We show that the length and level of expression of the transcript of CG13102 in testis are not affected in the mutant flies but we find that the 5' UTR of Rcd-1r in testis of mutant flies to be shorter than that of wild type flies. Western blot revealed that the amount of the retrogene protein is lower in sterile males was less than fertile males. Given these results we propose that Rcd-1r has a major role in the Drosophila male germline. We also studied Rcd-1r and Rcd-1 molecular evolution and reveal that the parental gene is under very strong purifying selection while the retrogene is evolving under both purifying and positive selection. McDonald-Kreitman test reveals that many amino acid substitutions have fixed not only to change function of this gene after duplication but after that in every lineage the protein keeps changing fast under positive selection. The evolutionary studies of Rcd-1r suggest that this gene may have a new function. This pattern of evolution is often seen in male germline genes that have functions that are under constant sexual selection or intersexual co-evolution. This would be different from the parental gene function.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS... ii

ABSTRACT iii

LIST OF ILLUSTRATIONS... vii

LIST OF TABLES... viii

Chapter

1. INTRODUCTION………………………………………………………. 1

1.1 Retroposition as a gene duplication mechanism……………………… 1

1.2 Required for cell differentiation 1(Rcd-1)…………………………….. 2

1.3 Rcd-1 related.. 5

1.4 Stock 11773………………………………………………………….. 7

1.5 Goals and summary of this work……………………………………… 9

2. MATERIALSAND METHODS…………………………………………… 10

2.1 Strains used…………………………………………………………… 10

2.2 Transcript analyses……………………………………………………. 11

2.3 Protein analyses……………………………………………………….. 12

2.4 Sequence Analyses…………………………………………………… 13

3. RESULTS ………………………………………………………………... 15

3.1 Checking mutant stock 11773... 15

3.2 5'RACE……………………………………………………………… 17

3.3 3'RACE……………………………………………………………….. 19

3.4. Quantitative PCR... 21

3.5. Protein analyses of *Rcd-1* and *Rcd-1r* in mutant stock (11773)… 24

3.6. Sequence analyses………………………………………………… 26

4. DISCUSSION…………………………………………………………….. .. 31

APPENDIX

A. ALIGNMENT OF *Rcd-1r* WITH PARENT GENE Rcd-1……………….. 35

B. PRIMER SEQUENCES AND THEIR USE………………………………. 40

C. Ka/Ks VALUES OF DIFFERENT PAML MODELS…………………….. 43

D. QUANTITATIVE RT_PCR Ct VALUES FOR EVERY
GENE AND REPLICATES…………………………………………………… 45

E. ALIGNMENT OF Rcd-1 CARBOXY TERMINUS WITH
HUMAN Rcd-1 REGION USED TO RAISE ANTIBODIES………………… 47

REFERENCES………………………………………………………………… 49

BIOGRAPHICAL INFORMATION………………………………………… 52

LIST OF ILLUSTRATIONS

Figure Page

1.1 Illustration of the process of retroposition... 2

1.2 Alignment Rcd-1 in different species... 4

1.3 Phylogenetic distribution of *Rcd-1r* in *Drosophila*……………………………… 6

1.4 Organization of the region where *Rcd-1r* is present……………………………… 7

1.5 P element insertion site shown in green………………………………………. 8

3.1 Position and orientation of the P element ……................................... ………… 16

3.2. Transcript Analysis of *CG13102* in wild type and sterile males……………… 18

3.3 Transcript Analysis of *Rcd-1r* in wild type and sterile males………………… 19

3.4 Results from 5'RACE ………………………………………………………. 20

3.5 Results of 3'RACE ………………………………………………………….. 21

3.6. Primers designed for quantitative PCR ……………………………………….. 22

3.7. The results from ANOVA…………………………………………………… ... 23

3.8. Results from Western blot………………………………………………………… 26

3.9 Phylogenetic Analysis of Rcd-1 and Rcd-1r…………………………………… 27

3.10. Comparison of two ratio model estimate with two ratio model fixed. 28

LIST OF TABLES

Table Page

3.1 Results of Quantitative PCR 24

3.2 Results of Mac Donald Kreitman Test………………………………………. 30

CHAPTER 1

INTRODUCTION

1.1 Retroposition as a gene duplication mechanism

Retroposition is a molecular process that leads to the formation of intronless gene duplicates (i.e. retrogenes). Retroposition occurs when the parental gene mRNA is reverse transcribed by the reverse transcriptase enzyme of the non LTR retrotransposable elements giving rise to a complementary DNA that gets randomly inserted in the genome (Esnault, Maestre et al. 2000). See figure 1.1 for an illustration of the process. Retrogenes are characterized by the absence of introns, presence of direct repeats and poly-A tail but these last two features are often not found in old retrogenes (Betrán, Thornton et al. 2002). These retroposed gene duplicates may either accumulate mutations resulting in the formation of functionless retropseudogenes or may become functional retrogenes. The functional retrogenes will have to recruit regulatory regions to be expressed (Betrán, Thornton et al. 2002).

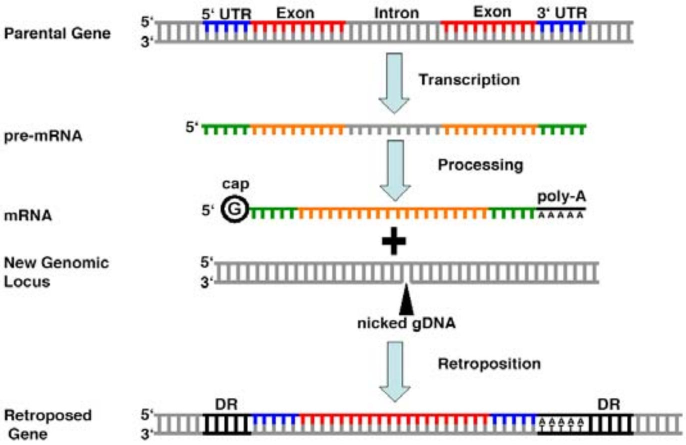


Figure: 1.1 Illustration of the process of retroposition. The parental gene is transcribed and an mRNA forms that is then converted to cDNA by reverse transcriptase and gets randomly inserted in the genome.

1.2 Required for cell differentiation 1 (*Rcd-1*)

Rcd-1 stands for *Required in cell differentiation 1*. It was Garces et al. (Garces, Gillon et al. 2007) who first identified the *Drosophila Rcd-1* (also known as *CG14213*). However, all the functional and structural studies for this gene have been carried out in yeast and in human. *Saccaromyces cerevisiae Rcd-1* is responsible for the nitrogen starvation induced cell differentiation (Tsukahara et al.1998). It was reported by Garces et al. that the homolog of *Rcd-1* in humans functions as a cell differentiation co-factor (Garces, Gillon et al. 2007). This gene has been shown to react with c-myb, a transcriptional factor required for haemopoetic differentiation in humans (Haas et al. 2004). This interaction was studied with the help of yeast two-hybrid assay. Rcd-1 consists of 18σ helices that give rise to the armadillo like repeats (Garces, Gillon et al.

2007). These repeats that are known to play a role in protein-protein interaction, are believed to help in the dimerization of Rcd-1. (Garces, Gillon et al. 2007). Rcd-1 dimer has recently been described to have a DNA binding cleft formed of positively charged amino acids that provides DNA binding affinity (Garces, Gillon et al. 2007). Rcd-1 is present and quite conserved in all eukaryotes. For example, Rcd-1 in humans is 62.5% identical to Rcd-1 in *Drosophila*. The comparison of Rcd-1 in several eukaryotes is shown in figure 1.2.

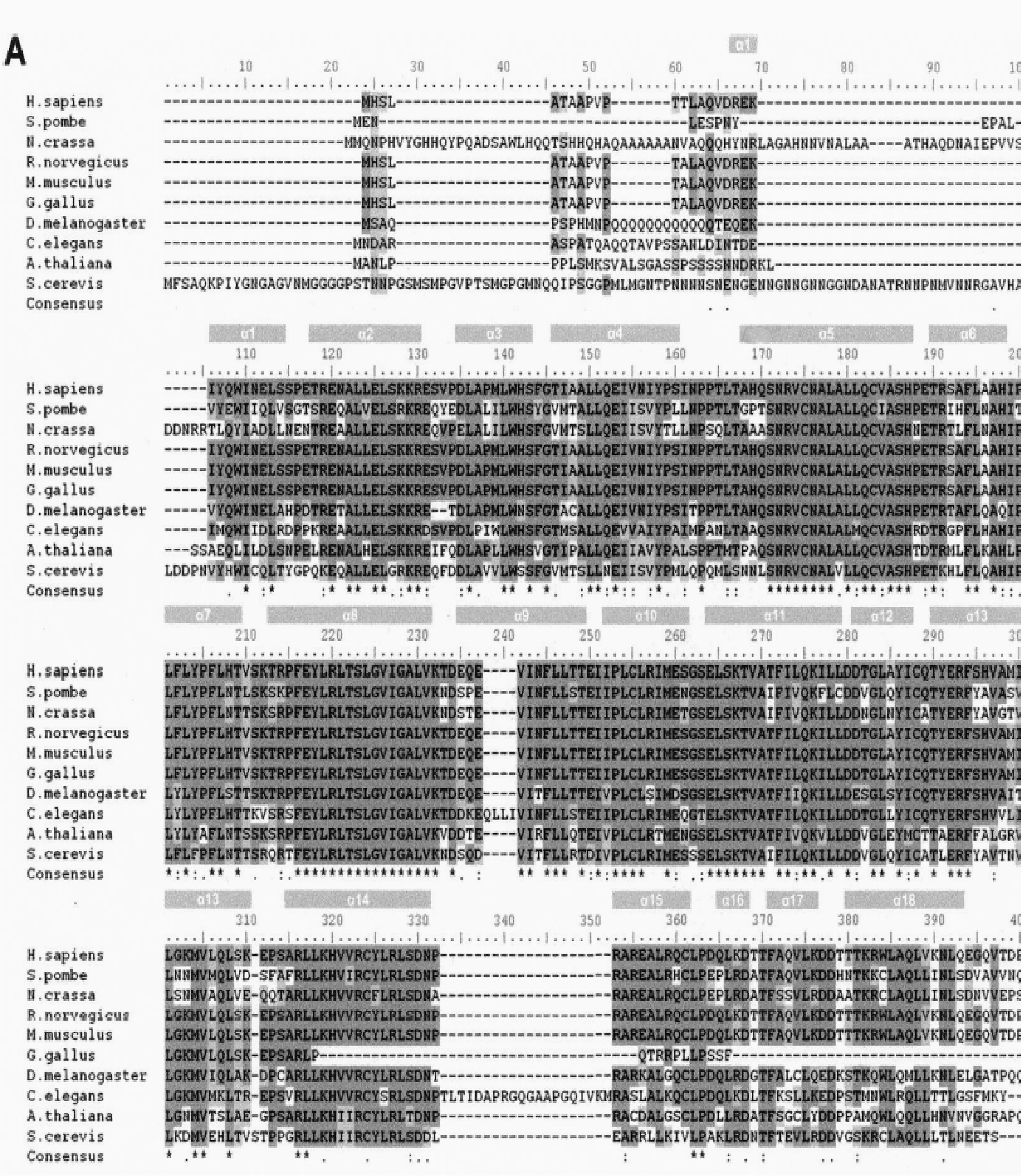


Figure: 1.2 Alignment of Rcd-1 in different species (From Garces *et al.* 2007*).*

1.3 *Rcd-1* related

In this work we study *Rcd-1 related* (*Rcd-1r*), a retrogene that originated from the parental gene *Rcd-1* through the process of retroposition. *Rcd-1r* is located in region 29 of the 2L chromosomal arm and present only in *Drosophila melanogaster*, *D. simulans*, *D. sechellia* and *D. mauritiana*, whereas the parental gene is present in all 12 sequenced species of *Drosophila*. This suggests that *Rcd-1r* originated about 5-13 million years ago (Bai, Casola et al. 2007). Figure 1.3 shows when the retroposition of *Rcd-1r* has probably occurred during *Drosophila* evolution.

The parental gene *Rcd-1* resides on the X chromosome. Therefore, *Rcd-1r* shows the movement from the X chromosome to an autosome that has been described before for many retrogenes (Betrán, Thornton et al. 2002; Emerson, Kaessmann et al. 2004; Bai, Casola et al. 2007) many of which may acquire male germline function (Vinckenbosch, Dupanloup et al. 2006; Bai, Casola et al. 2007).

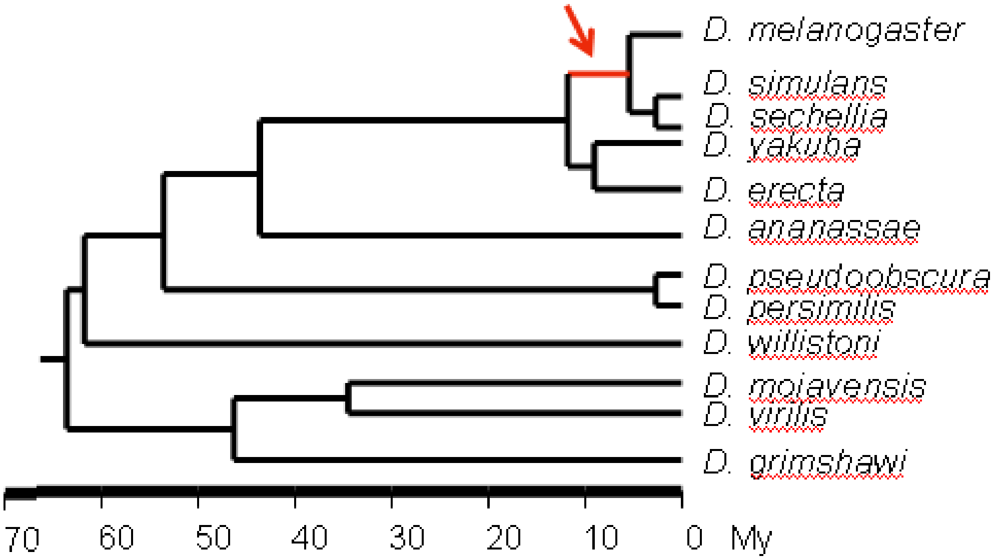


Figure: 1.3 Phylogenetic distribution of *Rcd-1r* in *Drosophila*. *Rcd-1r* origination (redrawn from (Bai, Casola et al. 2007) is shown in the *Drosophila* phylogeny.

In the process of retroposition *Rcd-1r* lost the 4 introns that are present in *Rcd-1*. An alignment of both genes from *D. melanogaster* is shown in the appendix. The genes show only 61.04 % identity in *D. melanogaster* at the nucleotide level. Interestingly, *Rcd-1r* inserted in the 3'UTR of another gene, *CG13102*, and in the complementary strand. Figure 1.4 shows the location of *Rcd-1r* relative to *CG13102*. *CG13102* is present in all 12 species of *Drosophila* examined, however its function is unknown.

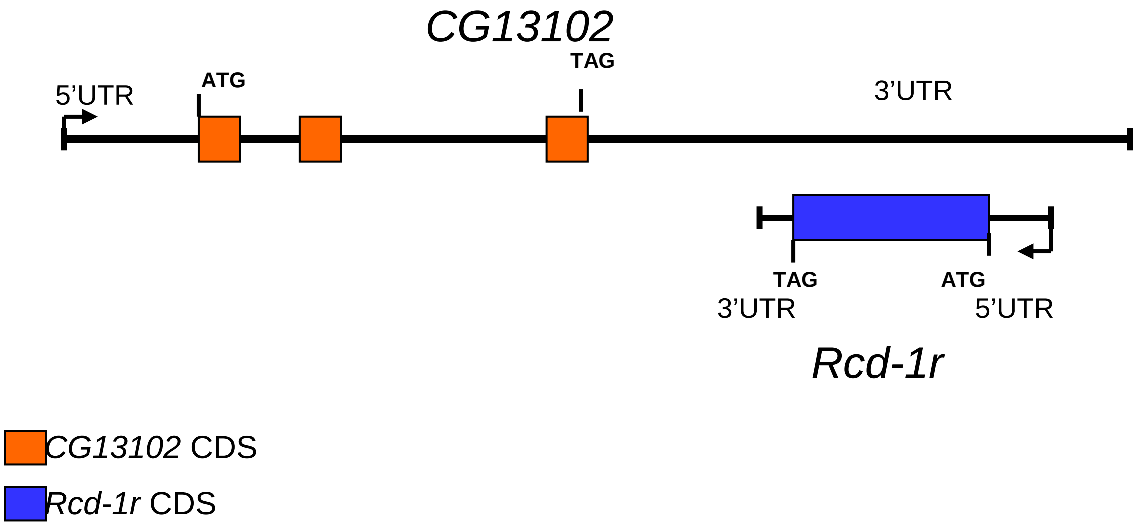


Figure: 1.4 Position of *Rcd-1r* relative to *CG13102.* The annotation of the 3'UTR of *CG13102* is based on the data available in fly base. The coding region of *CG13102* and *Rcd-1r* are shown in orange and blue respectively. The arrows in this figure indicate the transcription start sites.

1.4 Stock 11773

Castrillion *et al.* (Castrillon, Gonczy et al. 1993) produced 83 recessive autosomal mutants in *Drosophila* using P element insertions. Since the mutation induced by the P element was recessive, the phenotype observed in the mutant strain could be confirmed by crossing the strain homozygous for the P element insertion with a deletion line spanning the region of insertion of P element. Castrillion and collaborators thus crossed mutant lines homozygous for P element insertions with strains having deletions spanning the region of insertion of P element and confirmed the mutant phenotype. One of these mutants (stock 11773) had a P element inserted in the

5’UTR of *Rcd-1r* (i.e. 3’UTR of *CG13102*). Males homozygous for this P element insertion were sterile. Figure 1.5shows the site of insertion of the P element relative to *Rcd-1r* and *CG13102*.

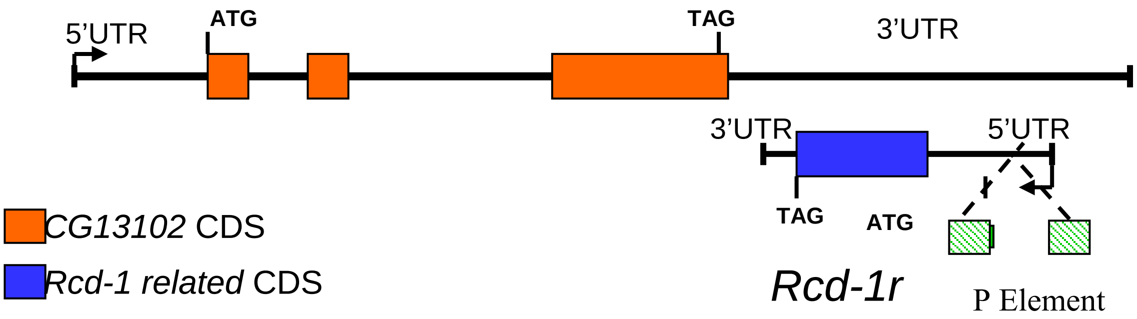


Figure: 1.5 P element insertion site with respect to *Rcd-1r* and *CG13102*.The P element is inserted in the 5’UTR of *Rcd-1r* in the complementary strand and 3’UTR of *CG13102*.

Castrillion *et al.* also crossed the stocks containing the P element insertion with another line providing the transposase enzyme that could excise the P element. They confirmed that the excision of the P element restored the fertility indicating that the sterility in the mutants was a result of the P element insertion. They also classified the stocks based on the stage of spermatogenesis that was altered as a result of the P element insertion. The stock 11773 having a P element insertion in the 5’UTR of *Rcd-1r* has been classified as post -meiotic differentiation defect mutant. These mutants undergo normal meiosis but the germline cells fail to differentiate into mature spermatids.

1.5 Goals and summary of results of this work

Rcd-1r is a retrogene of recent origin (Bai, Casola et al. 2007). In this work, we set up to study its function and mode of evolution. We address the function of the gene by trying to understand whether the P element inserted in its 5'UTR is knocking out the expression of *CG13102* or *Rcd-1r*, or both, and by studying the expression of both genes in wild type. We reveal that only *Rcd-1r* is testis specific and find that the 5'UTR of the *Rcd-1r* transcript in mutants is much shorter than the wild type males. Besides the total protein in mutant males is less than the wild type males. Amount of *Rcd-1r* protein as compared to the *Rcd-1* protein in mutant males is less than the wild type males. We conclude that the sterility phenotype is very likely due to the effect of the P element insertion on *Rcd-1r*. We use sequence analyses of polymorphism and divergence and reveal that Rcd-1r is evolving both under purifying and positive selection

CHAPTER 2

MATERIALS AND METHODS

2.1 Strains used

The expression of *CG13102* and *Rcd-1r* was first studied in wild type *D. melanogaster* flies from strain EC-180 (Ecuador-180). This stock was obtained from the Ballard lab. The stock (11773) containing the P element insertion (Castrillon, Gonczy et al. 1993) was obtained from Berkeley Drosophila Genome Center. The P element stock was balanced using CyO that is the balancer for the 2nd chromosome in *D. melanogaster*. The stock is mutant for the rosy eye color on the 3rd chromosome and this mutation is rescued by the P element insertion that contained the rosy gene. Because the P element insertion leads to recessive sterility in males, the stock is maintained with heterozygote males. After receiving the stock, it was checked to confirm the position of the P element insertion. DNA was extracted and PCRs performed with flanking primers and primers in the P element and flanking region. The genomic DNA extraction was carried out using the quick DNA isolation protocol (PUREGENE).

Isolines from the wild were used for our polymorphism analyses. Twelve isolines for *D. melanogaster* from single population from Zimbabwe and 10 for D. simulans from single population in Madagascar were used in these analyses. The Zimbabwe lines were kindly provided by the Wu lab. The Aquadro lab provided the Madagascar lines. A table with the names of the lines is shown in the Appendix.

2.2 Transcript analyses

RNA was extracted from whole adult males, testis and male carcass (that is the gonactomised males) of *EC-180* and the mutant stock 11773. 100 testes were dissected and kept in RNA later. RNA was extracted according to the protocol in the QIAGEN kit. The RNA was quantified using nano drop. RNA was first subjected to DNase digestion to remove any genomic DNA contamination. RT-PCRs were then carried out using the primers shown in table A1 in the appendix. Forward and reverse primers were designed for *CG13102* in the coding region and two others were designed in the coding region of *Rcd-1r* that is the 3'UTR of *CG13102*. A fifth primer was designed in the 3'UTR of *Rcd-1r* referred to as the specific primer. This primer was used instead of oligo dT for the first strand cDNA synthesis in the *Rcd-1r* RT-PCR. Given that the transcripts of both genes are believed to completely overlap (figure 1) this primer used instead of oligo dT allows for the study of the transcrption of *Rcd-1r*. *Gapdh2* was used as a positive control of the oligo dT RT. The PCR products were purified using QIAGEN PCR purification kit and sequenced using ABI sequencer and fluorescent DyeDeoxy terminator reagents.

5' and 3' RACE were performed in RNA from testis or fly halves of wild type males and males from the stock 11773 to reveal if the P element insertion affects transcriptional start and end sites. RACE stands for rapid amplification of complementary DNA ends. 5'RACE was done to identify the transcriptional start site of *Rcd-1r* and 3' RACE was performed to identify the transcription end site of *CG13102*.

We used Ambion kit 1700 for both 5’ and 3’RACE. Primer sequences are given in Appendix. PCR products were extracted from the gel and sequenced.

Quantitative real time-PCR (QRT-PCR) was used to quantify amount of RNA produced in the mutant line. RNA was extracted from fly halves of curly males heterozygous for the P element insertion and non-curly males homozygous for P element insertion. The two sets of primers one set in the coding region of *CG13102* and the other set in the overlapping region shared by both *CG13102* and *Rcd-1r* were used. The two sets of primers one set in the coding region of *CG13102* and the other set in the overlapping region shared by both *CG13102* and *Rcd-1r* were used. *Gapdh2* was used as control. Promega kit was used for these reactions and run in a 7300 ABI QRT-PCR machine. Primers used are shown in table A1 of the appendix. Products were run in a gel to control for spurious amplification.

2.3 Protein analyses

Western blots were performed for tissues of stock 11773. 80 testes of non-curly males, 80 testes of curly males and 35 fly female heads from curly and non curly females were separately mashed up in extraction buffer which was then incubated for 30 minutes at 4◦C. The samples were then spun at 13000rpm for 15 minutes and the supernatant transferred to a new eppendorf tube. The amount of protein was then quantified using Bradford reagent and equal amount of protein for each tissue sample was loaded on to a Criterion gradient gel (Biorad) and blotted on to nitrocellulose membrane that was probed with mouse polyclonal antibodies raised against residues

199-298 of human Rcd-1 (Abcam). All other steps of the western blot were performed according to the protocol in Current protocols in molecular biology (Ausubel, Brent et al. 1988). The alignment of the region of the human Rcd-1 used to raise the mouse polyclonal antibody with fly Rcd-1 and Rcd-1r is shown in the figure A 4 of the appendix. The carboxy terminus of Rcd-1 and Rcd-1r show 69% and 56.5% identity to the human antigenic region used respectively.

2.4 Sequence Analyses

The Ka/Ks ratio (ratio of nonsynonymous substitution per nonsynonymous site to the synonymous substitution per synonymous site) for both the parental gene *Rcd-1* and its retrogene *Rcd-1r* were compared using the PAML3.1 software (Yang 1997) and sequences from several of the *Drosophila* sequenced genomes (see below (Clark, Eisen et al. 2007). For this analysis a tree was provide that is shown in results section. This tree was constructed taking into account the phylogenetic relationships between species (Ting, Tsaur et al. 2000) (Clark, Eisen et al. 2007) and the age of the genes (Bai, Casola et al. 2007). Several models that made sense a priory were compared. We studied if retrogene and parental gene evolve at the same rate. We also address if the retrogene is now evolving under positive or under purifying selection. This was done by comparing a model with single rate, with a model with two rates, and other models in which these rates are fixed to 1. The comparisons were tested considering that twice the difference of likelihood between the models that we want to compare should distribute as a χ^2 with as many degree of freedom as the difference in parameters between the models (Yang 1997).

The McDonald-Kreitman test (McDonald and Kreitman 1991) was used to compare the within species polymorphism and between species divergence and further understand the mode of evolution of *Rcd-1r*. This test that contrasts the ratio of the fixed synonymous differences to the nonsynonymous differences with the ratio of the polymorphic synonymous to nonsynonymous sites was performed using *D. melanogaster* and *D. simulans* data. It was performed using DNAsp software (Rozas, Sanchez-DelBarrio et al. 2003). Isolines for *D. melanogaster* from Zimbabwe and for *D. simulans* from Madagascar were used for this analysis (see above). DNA was extracted for single fly for each of *D. melanogaster* and *D. simulans* isolines. DNA extractions were carried out using the Puregene kit with modification for single fly.

CHAPTER 3

RESULTS

As commented in the introduction, a P element inserted in the 5' UTR of *Rcd-1r* (that it is also annotated as the 3' UTR of *CG13102*) results in male sterility. Because of this we are trying to figure out which one of the two genes is responsible for the infertility phenotype and their possible role in spermatogenesis. Not much was known about the expression or transcript length of *CG13102* and *Rcd-1r* in males before this work. If fact the annotation of the *CG13102* transcript in FlyBase shown in figures 3.2 and 3.3 was supported by an embryo cDNA and the one for *Rcd-1r* by a cDNA from testis. Therefore we carried out transcription analyses of both genes in adult males in a wild type strain (*EC180*). We also carried out expression analyses in the mutant line. We began by confirming the P element insertion site and the orientation of P element in sterile males of stock 11773.

3.1 Checking mutant stock 11773

Stock 11773 was checked for the insertion of a P element in the described position (Castrillon, Gonczy et al. 1993). The stock is recessive male sterile and is maintained against the CyO balancer stock. Using the primers flanking the described site of insertion of P element we confirm the position of the P element. We did not get any PCR product for the non-curly flies (i.e. homozygous for the P element insertion) genomic DNA but got a PCR product for the curly flies of the stock (see Figure 3.1). Afterwards, we confirmed the orientation of the P element using one primer in the flanking region and one primer in the 5’ end of the P element. The PCR product is shown in figure 3.1. These PCR products were sequenced and corresponded to the expected sequences.

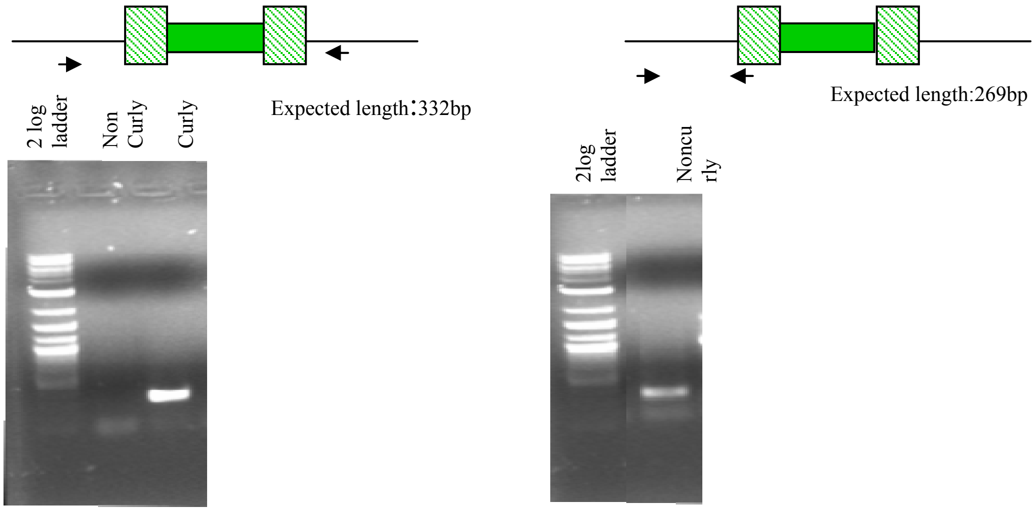


Figure: 3.1 Position and orientation of the P element and PCR products for 11773 strain checking. Primers were designed in the flanking region of the P element to confirm the P element insertion site. Primers were designed in the 5’end of the p element and the flanking region to determine the orientation of the P element.

11773 strain was also checked for the expected recessive male infertility phenotype. Different crosses were carried out to check the fertility of both the curly and non-curly flies. The curly and non-curly male flies were crossed separately to wild type (EC-180), and curly and non-curly virgin females flies from 11773. Two replicates were performed. Only the crosses involving non-curly males failed to produce offspring. Thus we confirmed that the non curly males were sterile whereas the non curly females, curly males and curly females were fertile.

3.2 RT-PCRs for *CG13102* and *Rcd-1r*

RT-PCRs were carried out for *CG13102* and *Rcd-1r* in testis and the rest of the body (i.e. carcass). A specific primer instead of oligo dT was used to make the cDNA for *Rcd-1r* to make sure we got amplification only when this gene was transcribed (see materials and methods). Figure 3.2 and 3.3 depict the expression of *CG13102* and *Rcd-1r* in testis and in carcass of wild type flies and testis of sterile flies. We observe transcription of both genes in testis but only *CG13102* is transcribed in carcass. For the RT-PCR from testis we used a positive control which was *Gapdh2*, a gene highly expressed in *Drosophila*. We also had RT negative for RNA sample of each gene. This RT negative lacked the superscript required for reverse transcription of mRNA into complementary DNA. This RT negative was very important to rule out any source of genomic contamination especially important for studying intronless retrogenes. Each RNA sample also had a PCR negative tube. This tube lacked the template from the RT. This tube was used to rule out any possible sources of contamination in the chemicals of the PCR reaction. We conclude that *Rcd-1r* is the only testis specific gene of the two and that both genes are still expressed in infertile males. We wanted to know if the transcripts are of different length in mutant males compared to wild type and performed 5’ and 3’ Race for *Rcd-1r* and *CG13102* respectively.

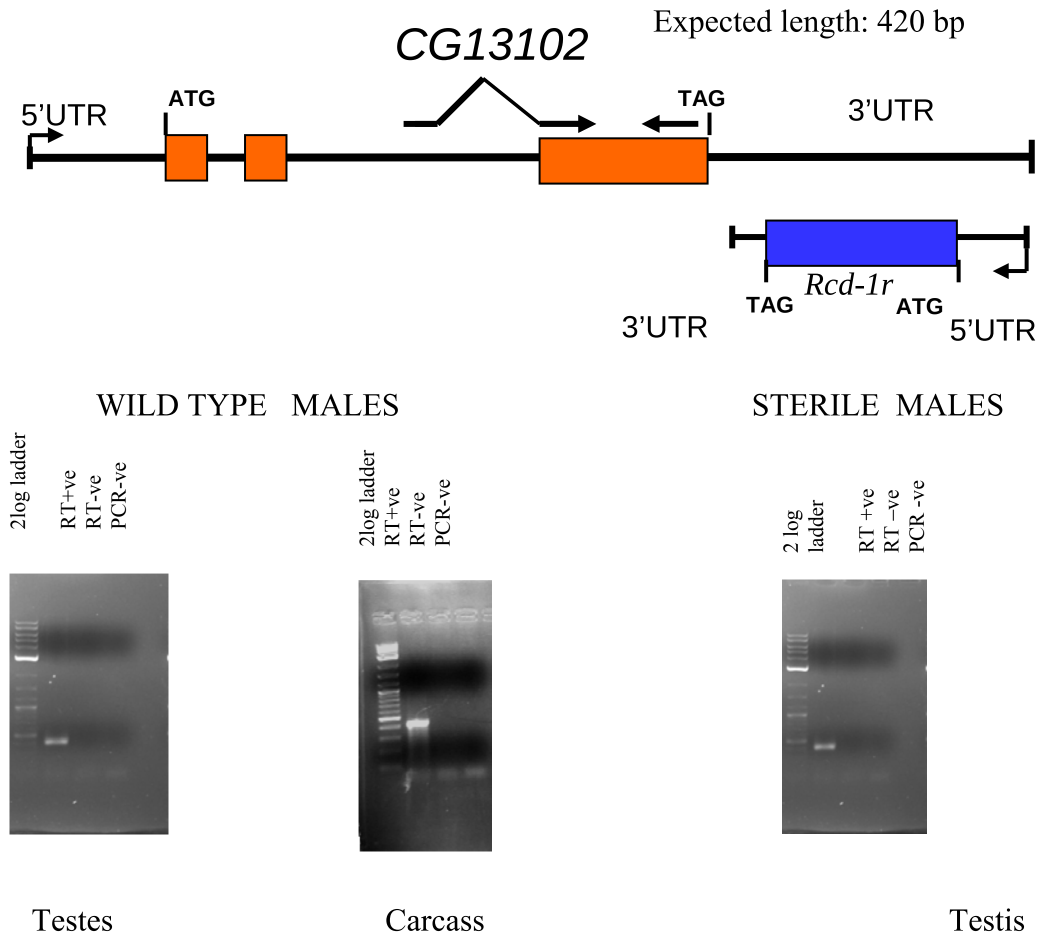


Figure: 3.2 Transcript analysis of *CG13102* in wild type and sterile males.

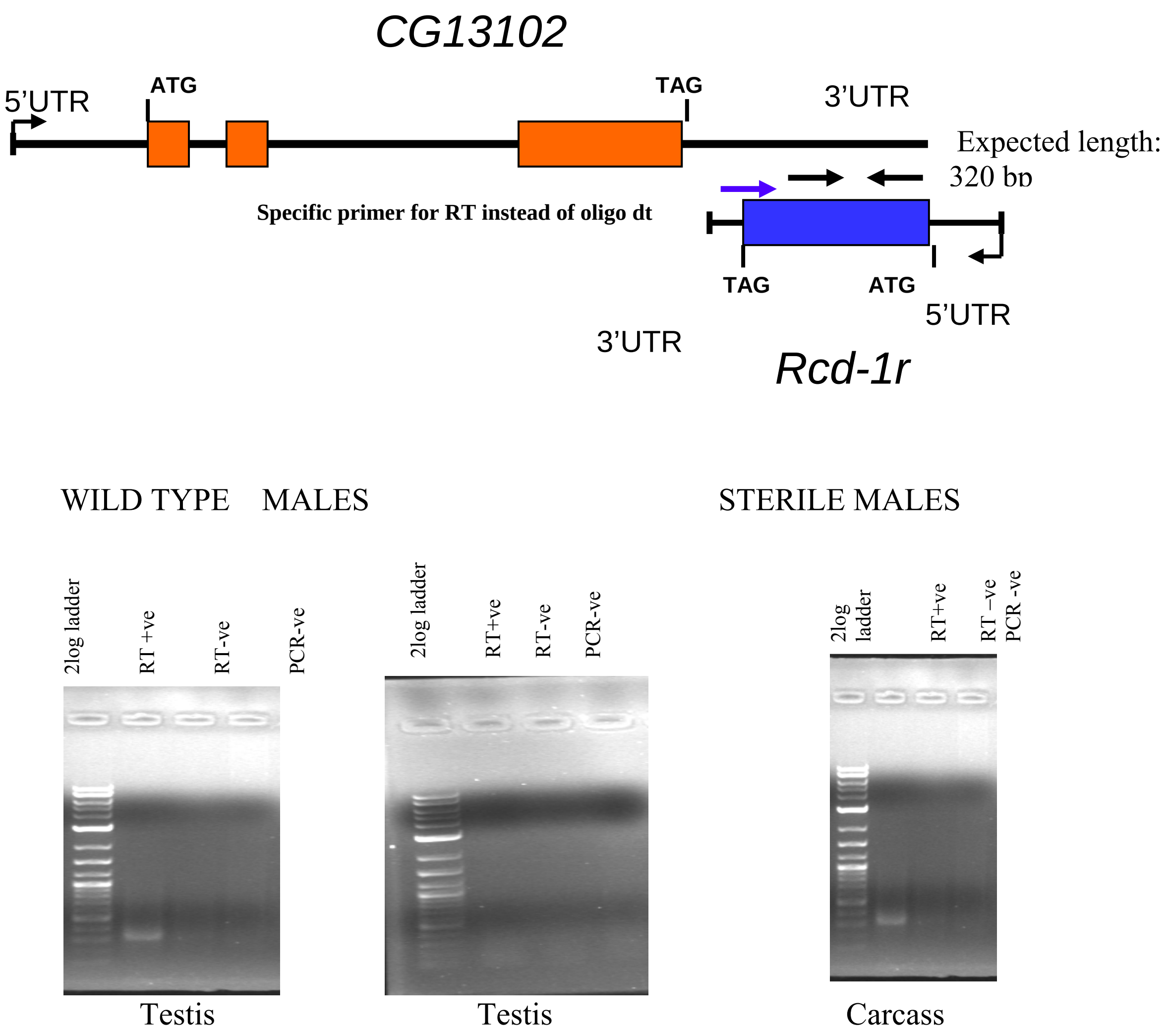


Figure: 3.3 Transcript analysis of *Rcd-1r* in wild type and sterile males.

3.3 RACE results

Afterwards, we studied if the P element insertion modifies the transcription start site (TSS) of *Rcd-1r* and the transcription end site (TES) of *CG13102* in wild type individuals of EC180???. We also performed a 5'RACE for the gene *Rcd-1r* and a 3'RACE for the gene *CG13102* using RNA from fly halves of non-curly males (i.e. mutant males). 5'RACE for the gene *Rcd-1r* was done in fly halves. Since we found the transcript of Rcd-1r to be testis specific, we used fly halves because in this type of RNA extraction the gene will be more represented than in RNA from the whole flies. The transcriptional start site of this gene was found to be the same as the one reported in the FlyBase annotation that was given by a testis cDNA of this gene. A diagram of the results of the 5' Race is depicted in figure 3.4. The PCR product obtained from 5'RACE for fly halves of non curly males was then sequenced to confirm length of the transcriptional start site of *Rcd-1r* in mutant males. The results indicated that the 5' UTR of *Rcd-1r* is shorter in non curly males than in wild type and it lies 116bp downstream of the transcriptional start site of wild type flies.

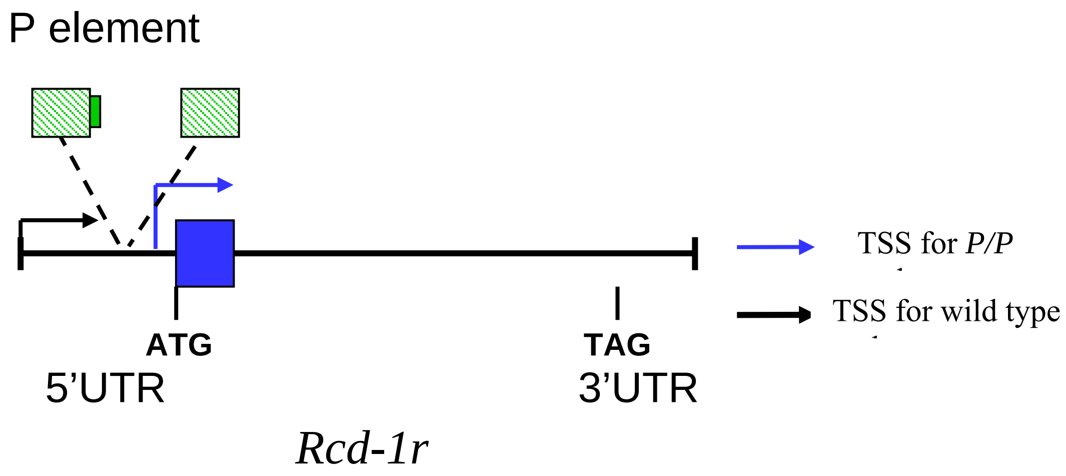


Figure: 3.4 Results from 5'RACE in *Rcd-1r*. The diagram above shows the length of *Rcd-1r* in wild type and mutant males (P/P).

3'Race product of *CG13102* was obtained from wild type testis. Interestingly, this product revealed that the 3'UTR of *CG13102* in embryo that was used to annotate the gene in FlyBase and reported to be 1308 bp in length is much shorter than the

3’UTR in testis. In testis, the 3’UTR was just 515 bp in length and did not extend till the P element insertion site. Revealing that the P element is unlikely to affect this gene transcript in this tissue. 3’Race product of *CG13102* obtained from testis of sterile males was the same length as in wild type (Figure 3.5).

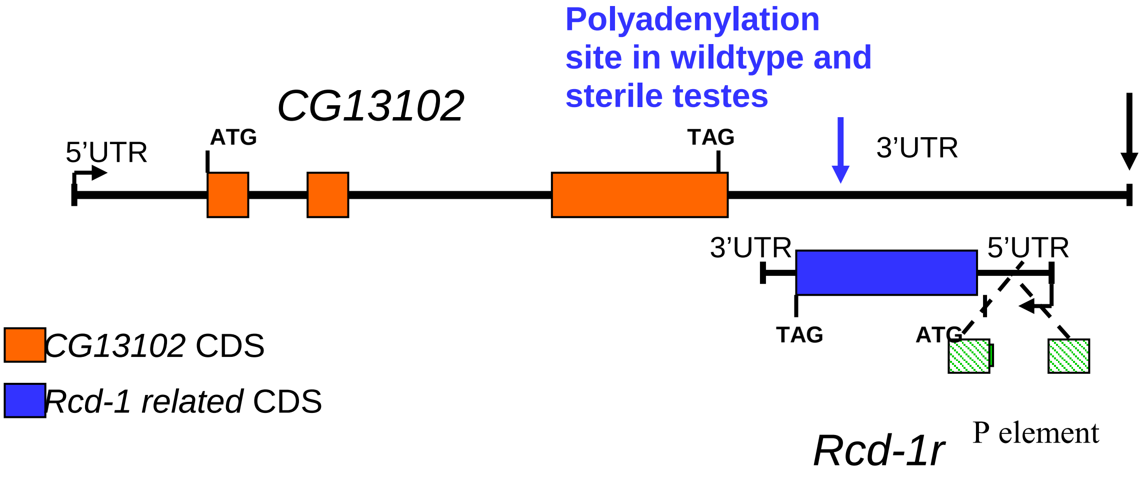


Figure: 3.5 Results of 3’RACE of *CG13102*. The figure shows the comparison of the 3’UTR as reported in embryo and compares the length of the same in wild type testes.

3.4. Quantitative real time PCR:

Given that both *Rcd-1r* and *CG13102* are still transcribed in mutant males, we wanted to quantify their level of transcription in comparison to heterozygote (Curly males). We carried out a quantitative real time PCR experiment with RNA from non-curly males and compared it to the RNA from the curly males. We analyzed 3 different RNA extractions from curly and non-curly flies and ran three different PCR reactions: (1) using the PCR primers 3 and 4 in the *Rcd-1r* region that overlaps *CG13102* (we call short product), (2) using primers 1 and 2 in the coding region of *CG13102* (we call long product) and (3) using primers for *Gapdh2* (normalizing control). The figure below shows the position of primers used for the PCR relative to *CG13102* and *Rcd-1r*.

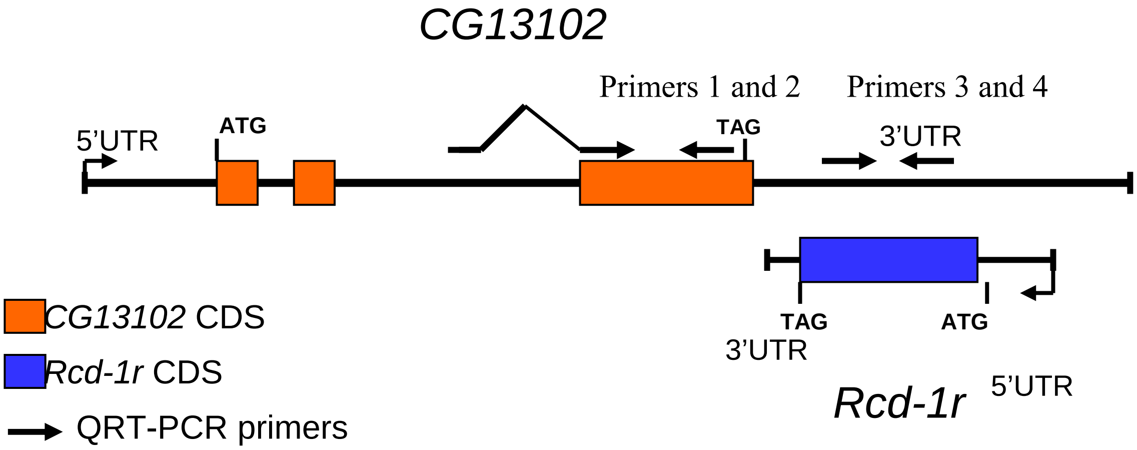


Figure: 3.6 Primers designed for quantitative PCR.The figure above indicates the primers obtained for quantitative PCR relative to *Rcd-1r* and *CG13102*.

The *Ct* values obtained from quantitative RT-PCR are given in a table in the appendix. We compared the ratio of short vs. long product and short and long vs. control using ANOVAs. Mean of *Ct* values for the ratio of short/long product in curly flies (fertile) was 1.454533 (SD=0.513106). Mean and the standard deviation for the *Ct* values obtained from the ratio of the short *vs.* long product in non-curly (infertile) males were 1.00633 and 0.110256, respectively. The $F_{(1,4)}$ was 2.1880 and we inferred no difference because P=0.2130.

Mean of *Ct* values for the ratio of long/control product in curly flies and non-curly flies were 0.8144 (S.D.=0.1299) and 1.0447 (S.D.=0.1181) respectively. The $F_{(1,4)}$ was 5.1587 and we inferred no difference because P=0.0850. The mean and the standard deviation of short normalized to *Gapdh2* for curly and non-curly males was 1.1554 (S.D.=0.3112) and 1.04466 (S.D.=0.11812). Therefore the ratios of short and long product normalized with *Gapdh2* did not differ between curly and non-curly males. The figure below illustrates the results obtained after comparing the short/GAPDH2 in curly and non curly males.

Table: 3. 1 Results from quantitative PCR

	Mean Ct Value	Standard Deviation
short/long-curly	1.454533	0.3112
short/long-non curly	1.00633	0.1181
long/gapdh2-curly	0.8144	0.1299
long/gapdh2-noncurly	1.0497	1.1181
short/gapdh2-curly	1.1554	0.31112
short/gapdh2-noncurly	1.04466	0.11812

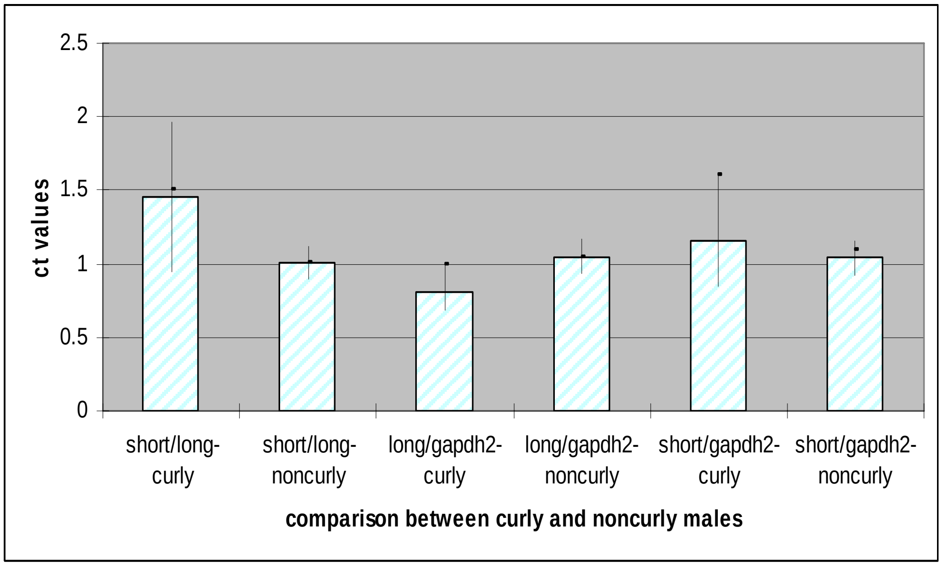


Figure: 3.7 QRT-PCR results. The table and the graph above depicts the results obtained by comparing the mean *Ct* values obtained for short and long product for curly and non curly flies.

Thus, first we infer that *Rcd-1r* is transcribed at very low levels because the ratio short/long is ~1. In addition, we conclude that there is no difference in the level of expression of either *CG13102* or *Rcd-1r* between curly and in non-curly males.

3.5 Protein analyses of *Rcd-1* and *Rcd-1r* in mutant stock (11773)

A western blot was performed for testes of non-curly males, testes of curly males and heads from a mix of curly and non-curly females. The results are shown in the figure below. The figure shows two bands in every lane. We interpret that as hybridization to both Rcd-1 and Rcd-1r in every tissue because according to their protein sequence they should be separated in a gradient gel by weight (i.e. 34 and 32 kilo Daltons respectively). We do not think that the two bands are posttranslational modifications of one of the two proteins because, previously, only a single band was observed for Rcd-1 westerns in rat (Hiroi, Ito et al. 2002). However, while we expected two bands in testis because is known from array analyses (http://www.flyatlas.org/) that both genes are expressed in testis, we expected only one strong band for the parental (i.e. the highest) in head. In head, array data has shown that *Rcd-1r* is down regulated and *Rcd-1* transcript is produced 18 times higher than the transcript for the retrogene. Therefore our interpretation is the Rcd-1r is produced from little transcript from head. These tissues were chosen because we wanted to see if Rcd-1r is not produced in testis of the sterile males (i.e. non-curly males). What we see is that despite loading the same amount of protein for the testes samples, we observe that the amount of Rcd-1r (lower band) protein in the testis of non-curly males appears to be less than the curly males. Rcd-1 also seems to be in fewer amounts given the intensity of band. This could be due to the effect of the P element (i.e. shorter 5' UTR) that could result in a diminished Rcd-1r protein production in the testis. However, several westerns need to be run to be able to quantify this. We also think that we might be able to raise an antibody for Rcd-1 to the region marked in green in figure A4 to help us confirm that the top band corresponds to Rcd-1.

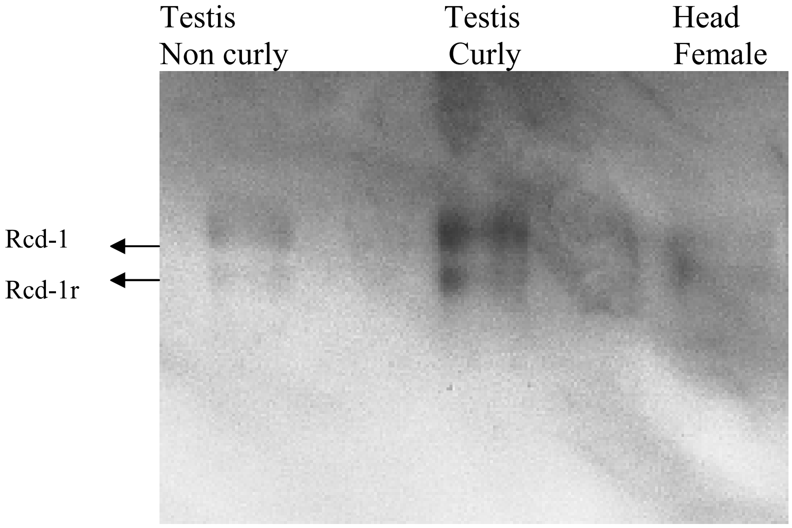
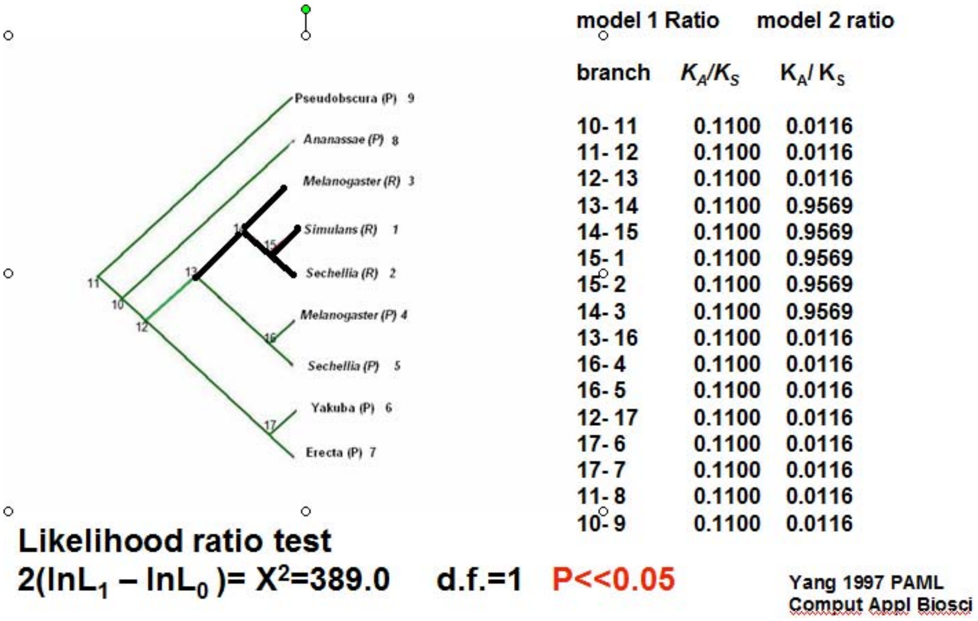


Figure: 3.8 Results from Western blot. The figure shows the bands obtained from western blot for testis of curly and non-curly flies and for head.

3.6. Sequence analyses

The mode of evolution of *Rcd-1r* and parental gene *Rcd-1* was analyzed using the PAML software (Yang 1997). The tree that was used with the sequences is shown in figure 3.9 and 3.10. Ka/Ks ratios for different branches were calculated under different models and their likelihoods compared. See table 2 in Appendix for all the details.

The free ratio model that assumes that all the branches evolve at different rates was run. The log likelihood value for this model was -3685.47113. The number of parameters for this model was 33. This model was then compared to the one ratio model that assumes that all the branches evolve at a single rate. The log likelihood value of one ratio model was -3685.47. The comparison revealed that free ratio model was significantly more likely than the one ratio model (X^2= 396.295, P<0.001, d.f.=15). This reveals that there are differences in rates of evolution in the different lineages. We then compared the two ratio that allows for different rate of evolution for the parent gene *Rcd-1* to that of the retrogene *Rcd-1r*. The two ratio model was significantly more likely than the one ratio model (X^2= 389.0 P<0.001, d.f.=1). The estimated rate of evolution of the retrogene was 0.9569.The figure below shows the comparison between the one ratio and the two ratio model.



	model 1 Ratio	model 2 ratio
branch	K_A/K_S	K_A/K_S
10- 11	0.1100	0.0116
11- 12	0.1100	0.0116
12- 13	0.1100	0.0116
13- 14	0.1100	0.9569
14- 15	0.1100	0.9569
15- 1	0.1100	0.9569
15- 2	0.1100	0.9569
14- 3	0.1100	0.9569
13- 16	0.1100	0.0116
16- 4	0.1100	0.0116
16- 5	0.1100	0.0116
12- 17	0.1100	0.0116
17- 6	0.1100	0.0116
17- 7	0.1100	0.0116
11- 8	0.1100	0.0116
10- 9	0.1100	0.0116

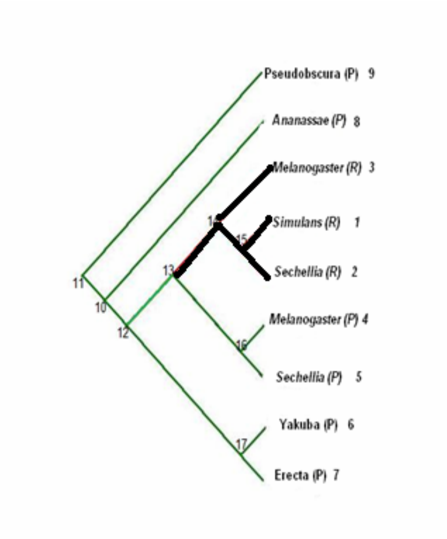
Likelihood ratio test

$2(\ln L_1 - \ln L_0) = X^2 = 389.0$ d.f.=1 P<<0.05

Yang 1997 PAML
Comput Appl Biosci

Figure: 3.9. Phylogenetic Analysis of Rcd-1 and Rcd-1r. The above phylogenetic tree shows the different parental and retrogene lineages. The retrogene lineages are marked by thick lines. The table shows the comparison of one ratio model with that of the two ratio model.

This is a very high rate and close to 1. We next test if this rate is different from 1. The two ratio model where the retrogene lineage was fixed to 1 and compared to the two ratio model. The comparison revealed that this two models do not significantly differ ($X^2 = 0.06$, $P<0.05$, d.f=1). This reveals that the retrogene is evolving at Ka/Ks ratio of ~1 (i.e. like a pseudogene). The results have been depicted using a phylogenetic tree below.



	model 2 Ratio ratio(fixed)	model 2
branch	K_A/K_S	K_A/K_S
10- 11	0.0116	0.0116
11- 12	0.0116	0.0116
12- 13	0.0116	0.0116
13- 14	0.9569	1.0000
14- 15	0.9569	1.0000
15- 1	0.9569	1.0000
15- 2	0.9569	1.0000
14- 3	0.9569	1.0000
13- 16	0.0116	0.0116
16- 4	0.0116	0.0116
16- 5	0.0116	0.0116
12- 17	0.0116	0.0116
17- 6	0.0116	0.0116
17- 7	0.0116	0.0116
11- 8	0.0116	0.0116
10- 9	0.0116	0.0116

Likelihood ratio test

$2(\ln L_1 - \ln L_0) = X^2 = 0.06$

d.f.=1 P>0.05

Yang 1997 PAML
Comput Appl Biosci

Figure: 3.10 Comparison of two ratio model estimate with two ratio model fixed.

The phylogenetic tree shows the comparison between two ratio models fixed to the model that estimates the rate of evolution between different lineages.

However, this does not necessarily imply that the *Rcd-1r* is a pseudogene because in all lineages the ORF is intact. It just points to the fact that is a very fast evolving gene. That is further explored below using the McDonald-Kreitman test.

In order to further explore the mode of evolution of *Rcd-1r* and reveal if the fast evolution is due to relaxation of constraint or positive selection we performed the McDonald-Kreitman test. This test compares the ratio of within species polymorphism for synonymous and nonsynonymous sites to the between species divergence. The table below shows the result obtained after comparing *D. melanogaster* with *D. simulans*. The ratio of the of fixed nonsynonymous to synonymous substitutions is 2.7576 and the ratio of the nonsynonymous to synonymous polymorphism is 0.3846 and the comparison is highly significant (see table 3.1). These ratios clearly show an excess of nonsynonymous fixed sites as compared to the nonsynonymous polymorphic sites which clearly shows that the protein is under positive selection between species. The number of polymorphic sites show an excess of synonymous sites as compared to the nonsynonymous sites which clearly proves that the retrogene *Rcd-1r* is also under purifying selection within species (see Table 3.1).

Table: 3.2 Results of Mc Donald Kreitman test:McDonald-Kreitman test for *Rcd-1r* between *D. melanogaster* and *D. simulans*

	Fixed	Polymorphic
Synonymous	33	13
Non synonymous	91	5

Fixed Non Synonomous/ Fixed Synonymous = 2.7576 Polymorphic non synonymous/ Polymorphic synonymous = 0.3846

X^2 =14.930, P=0.00011

CHAPTER 4

DISCUSSION

Examples are beginning to accumulate about new duplicates (often retrogenes) acquiring male germline specific expression in *Drosophila* (Long and Langley 1993; Yuan, Miller et al. 1996; Betrán, Thornton et al. 2002; Betrán and Long 2003; Hwa, Zhu et al. 2004; Tripoli, D'Elia et al. 2005; Arguello, Chen et al. 2006; Bai, Casola et al. 2007; Sturgill, Zhang et al. 2007). Some of this duplicates have been proven to have important functions in male germline (Timakov and Zhang 2001; Kalamegham, Sturgill et al. 2007; Zhong and Belote 2007). A set of them are transcription factors specific of the male germline that are required for postmeiotic differentiation (Hiller, Lin et al. 2001; Hiller, Chen et al. 2004; Chen, Hiller et al. 2005). The data we provide in this work suggests that *Rcd-1r* is another example of this type of duplicate that might be essential for fertility in *Drosophila*.

In this work, we address the function of *Rcd-1r*, a retrogene that is inserted in the 3'UTR of *CG13102* in its complementary strand. First, we studied transcription of *CG13102* and *Rcd-1r* in wild type males that clearly indicated that *Rcd-1r* is a testis specific gene whereas *CG13102* is transcribed throughout the body in *D. melanogaster*. We then compared the transcription level of *Rcd-1r* and *CG13102* in curly and non-curly males that had a P element inserted in the 5'UTR of *Rcd-1r* or the 3'UTR of *CG13102*. Contrary to our hypothesis we did not find differences in the level of expression between mutant and non-mutant flies.

However, when we studied the polyadenylation site of *CG13102* in wild type males and in sterile males which had a P element inserted in the 3'UTR, we inferred that this gene should not be affected by the P element. Although previous studies using a cDNA from embryo annotated the 3'UTR of this gene to be longer, we found that the transcript of this gene in wild type flies did not extend to the region of insertion of the P element. Besides there was no difference in the length of the 3'UTR of the sterile males as compared to the wild type males. These data supports that the P element does not affect *CG13102* transcript.

We then studied the transcription start site of both *Rcd-1*r and *CG13102* in wild type males and compared it to the sterile males. We found out that the transcriptional start site of *Rcd-1r* was 516 bp downstream from the transcriptional start site in wild type flies. Since we know that the elements required for translational and transcriptional control of testis expressed genes often reside in the 5'UTR of genes in *Drosophila* (Schafer, Kuhn et al. 1990; Kempe, Muhs et al. 1993; Schafer, Nayernia et al. 1995; Blumer, Schreiter et al. 2002), we hypothesize that *Rcd-1r* might have lost some control elements that were required for its transcription or translation. Since we know the transcript is made anda t the same level, we believe this could result in the following: 1. *Rcd-1r* transcript being made at the wrong time during spermatogenesis, or 2. the shorter Rcd-1r transcript would not translate to protein. To investigate the second possibility we did a western blot analysis and found that the amount of total protein in the testis of non-curly flies was less than the curly flies. Less protein than needed (i.e. an hypomorph) should often also lead to a phenotype. We could thus prove that the P element inserted in its 5’UTR affected only the expression *Rcd-1r* and we infer that this is causing the sterility.

Knowing that Rcd-1r was likely having a major role in spermatogenesis, we wanted to study its mode of evolution to describe whether this gene was in the process of acquiring a new function or kept the parental function. Very interestingly, sequence analyses of polymorphism and divergence revealed that *Rcd-1r* is evolving under purifying and positive selection. McDonald-Kreitman test reveals that many amino acid substitutions have fixed not only to change function of this gene after duplication but after that in every lineage the protein keeps changing fast under positive selection. This pattern of evolution is often seen in male germline genes and it has been explained by sexual selection or intersexual coevolution (Proschel, Zhang et al. 2006).

The results of this study clearly showed that *Rcd-1r* is a functional gene that is evolving under strong positive and purifying selection. This gene has a major role in male germline of Drosophila, since we know that a P element inserted in the 5’UTR of this gene results in male specific sterility. We report that the P element inserted in the 5’UTR of this gene results in formation of shorter transcript of *Rcd-1r* in sterile males. Although the gene is still transcribed in male germline, the loss of 5’UTR might lead to the loss of a translational repressor and the transcript may be produced at the wrong times in spermatogenesis. We tried to check for conserved motifs in the 5‘UTR but did not find any conserved motif. At the same time the total protein levels of *Rcd-1r* in sterile males is much less than the wild type males. This could be an additional effect of the P element insertion that needs to be further quantified. These results suggest that the

Rcd-1r has a very important in male germline and knocking out this gene results in sterility.

We however need to design experiments to check whether the *Rcd-1r* protein is made at a wrong time in spermatogenesis of sterile males resulting in sterility. If, on the other hand, the level of protein is the cause of the phenotype, introducing the wild type *Rcd-1r* transcript in the sterile flies that are homozygous for the P element insertion should restore the fertility. This experiments will give us further clues of the type of mutation the P element produces and of the role of *Rcd-1r* in spermatogenesis.

APPENDIX A

ALIGNMENT OF *Rcd-1r* WITH PARENT GENE *Rcd-1*

```
Rcd-1               -------CGGTTCTGGCCACACTGTTCGGCACACAAC-ATAATCCTGGTTGTCGCCACTT 52
mRNA Rcd-1          -------CGGTTCTGGCCACACTGTTCGGCACACAAC-ATAATCCTGGTTGTCGCCACTT 52
cds rcd-1           ------------------------------------------------------------
rcd-1r              ------------------------------------------------------------
cdsrcd-1r           ------------------------------------------------------------


Rcd-1               GTTTTCGATCATTTTAGACTTA-------AACCAATGCAAAAGGAATTTCC----TGCTA 101
mRNA Rcd-1          GTTTTCGATCATTTTAGACTTA-------AACCAATGCAAAAGGAATTTCC----TGCTA 101
cds rcd-1           ------------------------------------------------------------
rcd-1r              ------------------------------------------------------------
cdsrcd-1r           ------------------------------------------------------------


Rcd-1               ACGTTATCGATAATCGCAATAGACGCCAAATCGCCAGTTTTACCTCTTCCTTTTTCGCAG 161
mRNA Rcd-1          ACGTTATCGATAATCGCAATAGACGCCAAATCGCCAGTTTTACCTCTTCCTTTTTCGCAG 161
cds rcd-1           ------------------------------------------------------------
rcd-1r              ------------------------------------------------------------
cdsrcd-1r           ------------------------------------------------------------


Rcd-                GCCCCGTTCCGTTCAGCGAATTCTCGCCCGAGCTAAGTCTATTGCAAGCGTCCGGCTATA 221
mRNA Rcd-1          GCCCCGTTCCGTTCAGCGAATTCTCGCCCGAGCTAAG----------------------- 198
cds rcd-1           ------------------------------------------------------------
rcd-1r              ------------------------------------------------------------
cdsrcd-1r           ------------------------------------------------------------


Rcd-1               GAAATC--------GACGACAAATCGCTAGTTCCACCTCTTCCTTTTCGCAGGCCCCATT 273
mRNA Rcd-1          -----------------------------------------------------CCCCATT 205
cds rcd-1           ------------------------------------------------------------
rcd-1r              ------------------------------------------------------------
cdsrcd-1r           ------------------------------------------------------------


Rcd-1               CCGTTCAGCGAACTCTCGCCCGATCTAAGCGTCCGGCGATAGGAATC------------- 320
mRNA Rcd-1          CCGTTCAGCGAACTCTCGCCCGATCTAAGCGTCCGGCGATAGGAATC------------- 252
cds rcd-1           ------------------------------------------------------------
rcd-1r              ------GTCCTTTCTCGCTTGAAGCTTAT--ATCGATAACCACAATAGACGCCAACT--- 49
cdsrcd-1r           ------------------------------------------------------------


Rcd-1               ------CACCTCTTCCTTTTCGCAGGTCCCGTT----------CCGTTCAGCGAGTTCTC 364
mRNA Rcd-1          ------CACCTCTTCCTTTTCGCAGGTCCCGTT----------CCGTTCAGCGAGTTCTC 296
cds rcd-1           ------------------------------------------------------------
rcd-1r              ------CGAAACCACGTTTC---------------------------------------- 63
cdsrcd-1r           ------------------------------------------------------------


Rcd-1               GCCGACACTAAGTGCATTGCAAGCGTCCGGCGACAGTTCTTACCCAGATCTTGTCAAGTT 424
mRNA Rcd-1          GCCGACACTAAGTGCATTGCAAGCGTCCGGCGACAGTTCTTACCCAGATCTTGTCAAGTT 356
cds rcd-1           ------------------------------------------------------------
rcd-1r              ------------------TCAATCGTCCGTCGAAATTCCGGAGCCAGATCTTTCCAAGCT 105
cdsrcd-1r           ------------------------------------------------------------


Rcd-1               AACACG-------------------------------------ATGAGTGCTCAACCGAG 447
mRNA Rcd-1          AACACG-------------------------------------ATGAGTGCTCAACCGAG 379
cds rcd-1           -------------------------------------------ATGAGTGCTCAACCGAG 17
rcd-1r              ACTAGATCTTTGCAAGCTAACGCGAAAAAAAACCAGCTTTGCGATGAGTGCGGAACCAAG 165
cdsrcd-1r           -------------------------------------------ATGAGTGCGGAACCAAG 17
                                                              *******   * ** *

Rcd-1               TCCGCATATGAATCCTCAGCAGCAGCAGCAGCAGCAACAGCAGCAGCAGCAGACCGAGCA 507
mRNA Rcd-1          TCCGCATATGAATCCTCAGCAGCAGCAGCAGCAGCAACAGCAGCAGCAGCAGACCGAGCA 439
cds rcd-1           TCCGCATATGAATCCTCAGCAGCAGCAGCAGCAGCAACAGCAGCAGCAGCAGACCGAGCA 77
rcd-1r              TCCGGTAATGAGTCCCCAGCAGCAG---------------------------GCCGAGCG 198
cdsrcd-1r           TCCGGTAATGAGTCCCCAGCAGCAG---------------------------GCCGAGCG 50
                    **    ****  ** ***  ****                            ******

Rcd-1               GGAGAAGGTAAGCCGATGTCGGCAAGAACGCGTGCTAGCT--------------TTAACT 553
mRNA Rcd-1          GGAGAAGGT--------------------------------------------------- 448
cds rcd-1           GGAGAAGGT--------------------------------------------------- 86
rcd-1r              GGAGAAGGT--------------------------------------------------- 207
cdsrcd-1r           GGAGAAGGT--------------------------------------------------- 59
                   ******* *
```

```
Rcd-1               TATCGCCCGCCAGGTGTACCAGTGGATCAATGAGCTGGCCCATCCGGACACGCGTGAAAC 613
mRNA Rcd-1          ---------------GTACCAGTGGATCAATGAGCTGGCCCATCCGGACACGCGTGAAAC 493
cds rcd-1           ---------------GTACCAGTGGATCAATGAGCTGGCCCATCCGGACACGCGTGAAAC 131
rcd-1r              ---------------GTACCAGTTGATCATCGAGCTGGCCTATCCTGCCACGCGGGAGAC 252
cdsrcd-1r           ---------------GTACCAGTTGATCATCGAGCTGGCCTATCCTGCCACGCGGGAGAC 104
                                  ****** * *** *  *** ***** **** * * **** **

Rcd-1               CGCTCTGCTCGAGCTAAGCAAGAAGCGTGAGACGGACCTGGCCCCCATGCTGTGGAACAG 673
mRNA Rcd-1          CGCTCTGCTCGAGCTAAGCAAGAAGCGTGAGACGGACCTGGCCCCCATGCTGTGGAACAG 553
cds rcd-1           CGCTCTGCTCGAGCTAAGCAAGAAGCGTGAGACGGACCTGGCCCCCATGCTGTGGAACAG 191
rcd-1r              CGCTCTGCTGGAGCTGAGCAAGAACACCTATGCGGACCTGGCCCCCATGCTGTGGAAAAG 312
cdsrcd-1r           CGCTCTGCTGGAGCTGAGCAAGAACACCTATGCGGACCTGGCCCCCATGCTGTGGAAAAG 164
                    ********* ***** ********     *  * ** ******************** **

Rcd-1               CTTCGGGACCGCCTGCGCCCTGTTGCAGGAGATTGTTAACATCTACCCATCGATAACGCC 733
mRNA Rcd-1          CTTCGGGACCGCCTGCGCCCTGTTGCAGGAGATTGTTAACATCTACCCATCGATAACGCC 613
cds rcd-1           CTTCGGGACCGCCTGCGCCCTGTTGCAGGAGATTGTTAACATCTACCCATCGATAACGCC 251
rcd-1r              CGTCGGTACCACCTGCACCCTGCTGCAGGAGATCGTCAACATATACCCCATAATAACGAC 372
cdsrcd-1r           CGTCGGTACCACCTGCACCCTGCTGCAGGAGATCGTCAACATATACCCCATAATAACGAC 224
                      * **** ***   * * ***** ******** * ** *** * *****    *** **

Rcd-1               GCCCACTTTGACGGCCCACCAGTCGAACCGCGTGTGCAACGCCCTGGCGCTGCTGCAGTG 793
mRNA Rcd-1          GCCCACTTTGACGGCCCACCAGTCGAACCGCGTGTGCAACGCCCTGGCGCTGCTGCAGTG 673
cds rcd-1           GCCCACTTTGACGGCCCACCAGTCGAACCGCGTGTGCAACGCCCTGGCGCTGCTGCAGTG 311
rcd-1r              GCCCGTTTTGAAGGCCAACCAGTCGAACCGCGTGTGCTACGCCCTGACTTTGCTACAGTG 432
cdsrcd-1r           GCCCGTTTTGAAGGCCAACCAGTCGAACCGCGTGTGCTACGCCCTGACTTTGCTACAGTG 284
                     **    **** **** ******************** ******** *  * ** *****

Rcd-1               CGTCGCCTCGCACCCGGAGACTCGCACGGCCTTCCTGCAGGCCCAGATACCGCTGTACTT 853
mRNA Rcd-1          CGTCGCCTCGCACCCGGAGACTCGCACGGCCTTCCTGCAGGCCCAGATACCGCTGTACTT 733
cds rcd-1           CGTCGCCTCGCACCCGGAGACTCGCACGGCCTTCCTGCAGGCCCAGATACCGCTGTACTT 371
rcd-1r              CGTCGCCTCGCATCCGGAGACTCGCCCGGCCTTCCTGCGGGACCAGATACCGATGTACTT 492
cdsrcd-1r           CGTCGCCTCGCATCCGGAGACTCGCCCGGCCTTCCTGCGGGACCAGATACCGATGTACTT 344
                    ******* **** **** ******* ******** *** ** ********** * *** *

Rcd-1               GTACCCTTTCTTGTCGACCACGTCCAAGACCAGGCCCTTCGAGTACTTGCGCCTGACCAG 913
mRNA Rcd-1          GTACCCTTTCTTGTCGACCACGTCCAAGACCAGGCCCTTCGAGTACTTGCGCCTGACCAG 793
cds rcd-1           GTACCCTTTCTTGTCGACCACGTCCAAGACCAGGCCCTTCGAGTACTTGCGCCTGACCAG 431
rcd-1r              GTACCCCTTCTTGTCGACCACGTTCAAGAGCAGGCCCTTCGAGCAGCTGCGCCTGACCAC 552
cdsrcd-1r           GTACCCCTTCTTGTCGACCACGTTCAAGAGCAGGCCCTTCGAGCAGCTGCGCCTGACCAC 404
                     *** ** *** * ********** ***** ** ********** *  ********* **

Rcd-1               TTTGGGCGTGATTGGTGCTCTGGTCAAGGTAGGTTAAGCCGAAGGCCAGTGGTCATCTAA 973
mRNA Rcd-1          TTTGGGCGTGATTGGTGCTCTGGTCAAG-------------------------------- 821
cds rcd-1           TTTGGGCGTGATTGGTGCTCTGGTCAAG-------------------------------- 459
rcd-1r              GCTGGGCGTGATTAATGCTCTGGCCGAG-------------------------------- 580
cdsrcd-1r           GCTGGGCGTGATTAATGCTCTGGCCGAG-------------------------------- 432
                     ** **** ***  ******** * **

Rcd-1               TGCACTAATGCACTTGTTGTCTTCTTATTACTGCTATCGACAGACCGACGAACAGGAGGT 1033
mRNA Rcd-1          -------------------------------------------ACCGACGAACAGGAGGT 838
cds rcd-1           -------------------------------------------ACCGACGAACAGGAGGT 476
rcd-1r              -------------------------------------------ACCGGTGATACGGAGGT 597
cdsrcd-1r           -------------------------------------------ACCGGTGATACGGAGGT 449
                                                              ****  **    ****

Rcd-1               GATCACCTTTCTGCTGACCACCGAGATCGTGCCCCTTTGTCTGAGCATCATGGACAGCGG 1093
mRNA Rcd-1          GATCACCTTTCTGCTGACCACCGAGATCGTGCCCCTTTGTCTGAGCATCATGGACAGCGG 898
cds rcd-1           GATCACCTTTCTGCTGACCACCGAGATCGTGCCCCTTTGTCTGAGCATCATGGACAGCGG 536
rcd-1r              CCTCATCTTCCTGATATGGAGCGAGGTCGTGCCTCACTGTCTGACCAATATGGTCAGGGG 657
cdsrcd-1r           CCTCATCTTCCTGATATGGAGCGAGGTCGTGCCTCACTGTCTGACCAATATGGTCAGGGG 509
                      *** ***  *  *    * **** ******* *  ** ** * **  ***  **  **

Rcd-1               ATCGGAGCTGAGCAAGACTGTGGCCACTTTCATCATCCAGAAGATCCTGCTCGATGAGTC 1153
mRNA Rcd-1          ATCGGAGCTGAGCAAGACTGTGGCCACTTTCATCATCCAGAAGATCCTGCTCGATGAGTC 958
cds rcd-1           ATCGGAGCTGAGCAAGACTGTGGCCACTTTCATCATCCAGAAGATCCTGCTCGATGAGTC 596
rcd-1r              ATCGAAGCTGACCAAGATCGCGGCCACTTCAATCCTTGAGAAGATACTGCTCGATGAGAT 717
cdsrcd-1r           ATCGAAGCTGACCAAGATCGCGGCCACTTCAATCCTTGAGAAGATACTGCTCGATGAGAT 569
                    ****  *** * ****   *  ****     *** *  *******  ***** ** **

Rcd-1               GGGTCTGTCGTACATCTGCCAGACCTACGAACGCTTTTCGCACGTGGCCATCACCCTGGT 1213
mRNA Rcd-1          GGGTCTGTCGTACATCTGCCAGACCTACGAACGCTTTTCGCACGTGGCCATCACCCTGG- 1017
cds rcd-1           GGGTCTGTCGTACATCTGCCAGACCTACGAACGCTTTTCGCACGTGGCCATCACCCTGG- 655
rcd-1r              GGGTCTGACGTACATTTGCGAGAACCACGATCGCTTCTCGCAGGTGGCCATCACCCTGG- 776
cdsrcd-1r           GGGTCTGACGTACATTTGCGAGAACCACGATCGCTTCTCGCAGGTGGCCATCACCCTGG- 628
```

```
                   ******  ******* *** *** * **** *****  *    ******* *********

Rcd-1              GAGTGTCCGTTTCCCGAAGCCTTATGCATTGTCTAAGGGTCTTTATCGATTTCAGGGCAA 1273
mRNA Rcd-1         --------------------------------------------------------GCAA 1021
cds rcd-1          --------------------------------------------------------GCAA 659
rcd-1r             --------------------------------------------------------GCAA 780
cdsrcd-1r          --------------------------------------------------------GCAA 632
                                                                         ****

Rcd-1              AATGGTCATCCAGCTGGCGAAGGATCCATGCGCCCGGCTGTTGAAGCATGTGGTGCGCTG 1333
mRNA Rcd-1         AATGGTCATCCAGCTGGCGAAGGATCCATGCGCCCGGCTGTTGAAGCATGTGGTGCGCTG 1081
cds rcd-1          AATGGTCATCCAGCTGGCGAAGGATCCATGCGCCCGGCTGTTGAAGCATGTGGTGCGCTG 719
rcd-1r             AATGGTCATCCATATGTTGAAGTTTCCATGCCTCCGGGTGCTGAAGCATGTGGTGCGCTG 840
cdsrcd-1r          AATGGTCATCCATATGTTGAAGTTTCCATGCCTCCGGGTGCTGAAGCATGTGGTGCGCTG 692
                   ****  *** *   **  ***   *******  **** ** ****** ************

Rcd-1              CTATCTGCGTCTTTCGGACAATACACGGTAAGTCTTGGCGGGCGAGTAAATTATCGCTCG 1393
mRNA Rcd-1         CTATCTGCGTCTTTCGGACAATACACG--------------------------------- 1108
cds rcd-1          CTATCTGCGTCTTTCGGACAATACACG--------------------------------- 746
rcd-1r             CTATCTACTACTTACGGAAAATGCGCG--------------------------------- 867
cdsrcd-1r          CTATCTACTACTTACGGAAAATGCGCG--------------------------------- 719
                   ****** *   *  **** ***

Rcd-1              GCTTTTCAGTCTAGCCTTGACTCTTTCAGCGCTCGCAAGGCCCTGGGACAGTGTCTGCCC 1453
mRNA Rcd-1         -----------------------------CGCTCGCAAGGCCCTGGGACAGTGTCTGCCC 1139
cds rcd-1          -----------------------------CGCTCGCAAGGCCCTGGGACAGTGTCTGCCC 777
rcd-1r             -----------------------------CGCTCGCAGTGCACTTAGAGTGTGTCTGCCG 898
cdsrcd-1r          -----------------------------CGCTCGCAGTGCACTTAGAGTGTGTCTGCCG 750
                                               ** *****  ** **  **  *********

Rcd-1              GATCAGCTGCGTGACGGCACCTTTGCGCTGTGCCTGCAAGAGGACAAGTCGACCAAGCAG 1513
mRNA Rcd-1         GATCAGCTGCGTGACGGCACCTTTGCGCTGTGCCTGCAAGAGGACAAGTCGACCAAGCAG 1199
cds rcd-1          GATCAGCTGCGTGACGGCACCTTTGCGCTGTGCCTGCAAGAGGACAAGTCGACCAAGCAG 837
cdsrcd-1r          GATCTGCTGCGTGACGGCACCTTCACGTCGCTCGTGCAACATGACACGTGCACCAAGCAG 810
                  **** *** **************  *   *  * * *** * ***  **   ********

Rcd-1              TGGCTGCAGATGCTGCTCAAGAACCTGGAACTCGGCGCC------ACTCCGCAGCAGATC 1567
mRNA Rcd-1         TGGCTGCAGATGCTGCTCAAGAACCTGGAACTCGGCGCC------ACTCCGCAGCAGATC 1253
cds rcd-1          TGGCTGCAGATGCTGCTCAAGAACCTGGAACTCGGCGCC------ACTCCGCAGCAGATC 891
rcd-1r             TGGCTGCAGATGCTGCTCAAGAACCTGCAGACAAACGC---------------------C 997
cdsrcd-1r          TGGCTGCAGATGCTGCTCAAGAACCTGCAGACAAACGC---------------------C 849
                  ****** *** * ************** *       **

Rcd-1              GGCATGTCGCCACTGGGCTCCTAGGGAGTGGAAGCGACAGCAC-ACCCAGCACTATATTA 1626
mRNA Rcd-1         GGCATGTCGCCACTGGGCTCCTAGGGAGTGGAAGCGACAGCAC-ACCCAGCACTATATTA 1312
cds rcd-1          GGCATGTCGCCACTGGGCTCCTAG------------------------------------ 915
rcd-1r             GTCAACCCAATGGGCTCCTCCTAGAGA-----AGCGCC-----CAACAAC--TTTCGTCA 1045
cdsrcd-1r          GTCAACCCAATGGGCTCCTCCTAG------------------------------------ 873
                    * *    *         *******

Rcd-1              TCAATTCTAACGTCCAACCATCGATGATGAATAATCCACGCTGTTAATTACATAGCTGCT 1686
mRNA Rcd-1         TCAATTCTAACGTCCAACCATCGATGATGAATAATCCACGCTGTTAATTACATAGCTGCT 1372
cds rcd-1          ------------------------------------------------------------
rcd-1r             CTCCACCGTCCATTCAAGCGCCATCGATTAGTGATACCCACTCCTAGTTA----ACTGCT 1101
cdsrcd-1r          ------------------------------------------------------------


Rcd-1              CGACCGAATATACATAAACATG------CATACATACATACA------------------ 1722
mRNA Rcd-1         CGACCGAATATACATAAACATG------CATACATACATACA------------------ 1408
cds rcd-1          ------------------------------------------------------------
rcd-1r             CGACCGAATAAACATG-------------------------------------------- 1117
cdsrcd-1r          ------------------------------------------------------------


Rcd-1              --------------------------------------TACATATTTCATAAGCGTGCAC 1744
mRNA Rcd-1         --------------------------------------TACATATTTCATAAGCGTGCAC 1430
cds rcd-1          ------------------------------------------------------------
rcd-1r             ------------------------------------------------------------
cdsrcd-1r          ------------------------------------------------------------


Rcd-1              G-AGATTGCCTGTGTCGACTTAAGCGGAGCTGTAATATAC------CTACATAAATTACA 1797
mRNA Rcd-1         G-AGATTGCCTGTGTCGACTTAAGCGGAGCTGTAATATAC------CTACATAAATTACA 1483
cds rcd-1          ------------------------------------------------------------
rcd-1r             -------GCCT-------------------------------------ACATAATCTACA 1133
cdsrcd-1r          ------------------------------------------------------------
```

```
Rcd-1             AATAAATTAATGTTAATACTAAGATTGTCACTGACACCCAATAAAGAAC----------- 1846
mRNA Rcd-1        AATAAATTAATGTTAATACTAAGATTGTCACTGACACCCAATAAAGAAC----------- 1532
cds rcd-1          -----------------------------------------------------------
rcd-1r             AATGAAATTATGATATTACTAAAATTGACGGTA-----CAATAAAAAATATCAATAT---1178
cdsrcd-1r          -----------------------------------------------------------


Rcd-1              -------------------------------CTCGAGCAGACGATACGATAATA----- 1861
mRNA Rcd-1         -------------------------------CTCGAGCAGACGATACGATAATA------ 1547
cds rcd-1          -----------------------------------------------------------
rcd-1r             --------------------------------------- -------------------- 1186
cdsrcd-1r          -----------------------------------------------------------
```

Figure: A.1 Alignment of the parent gene *Rcd-1* and the retro gene *Rcd-1r* in *D. melanogaster*. The start and the stop codon of both the parent and the retrogene are marked in red.

APPENDIX B

PRIMER SEQUENCES AND THEIR USE

Table B. 1 PRIMER SEQUENCES AND THEIR USE

Gapdh2(forward)	5’-TCAGCCATCAGAGTCGATTC-3’	PCR
Gapdh2(reverse)	5-CAAACGAACATGGGAGCATC-3’	PCR
M_{13}(forward)	5-GTAAAACGACGGCCAG-3’	PCR& sequencing
M_{13}(reverse)	5’-CAGAAACAGCTATGAC-3’	PCR& sequencing
5’RACE(outer)for gene Rcd-1r	5’-TTGGCCTTCAAAACGGGCGT-3’	PCR& sequencing
5’RACE(inner) for gene rcd-1r	5’-TGCTGCTGGGGACTCATTAC-3’	PCR& sequencing
5’RACE kit(outer) for gene rcd-1r	5'-GCTGATGGCGATGAATGAACACTG	PCR& sequencing
5’RACE kit(outer)	5’CGGATCCGAACACTGCGTTTGCTGGCTTTGATG	PCR& sequencing
3’RACE kit(inner)	5'-GCGAGCACAGAATTAATACGACT-3'	PCR& sequencing
Primers in the overlapping(forward)	5’-CACATGCTTCAGCACCCGGA-3’	PCR& sequencing
Primers overlapping(reverse)	5’-AGCAGGCCCTTCGAGCAGCT-3’	PCR& sequencing
Primers coding region(forward)	5’-CAGCATACTTGTGGCCTTTG-3’	PCR& sequencing

Coding region primers(reverse)	5-CTTGACCAGCAAGAGACCCA-3’	PCR& sequencing
Specific primer	5’-AGCAGATCCGGCAGACACAC-3’	PCR& sequencing
MacDonaldkreitman(forward melanogaster)	5’-CACAATAGACGCCAACTCGAAACCAC-3’	PCR& sequencing
Kreitman test(reverse melanogaster)	5’-TGGCGCTTGAATGGACGGT-3’	PCR& sequencing
Kreitman test(forward simulans)	5’-CACAATCGACGCCAAATCGAATGCAC-3’	PCR& sequencing
Kreitman test(reverse simulans)	5’-ATGGCGCTTGCATGGAC-3’	PCR& sequencing
5’SUP	5’-TCCAGTCACAGCTTTGCAGC-3’	PCR& sequencing
3’SUP	5’-	
Flanking5’	5’-TTCCTCGGCTCTGGCCACAT-3’	PCR& sequencing
Flanking3’	5’-CCAGGTCCGCATAGGTGTTC-3’	PCR& sequencing
3’RACE(outer)	5’-AGCAGATCCGGCAGACACAC-3’	PCR& sequencing
3’RACE(inner)	5’-AGCAGGCCCTTCGAGCAGCT-3’	PCR& sequencing

APPENDIX C

Ka/ Ks VALUES OF DIFFERENT PAML MODELS

Table: C.1 Ka/ Ks VALUES OF DIFFERENT PAML MODELS

Branch	**Free ω**	**One ω**	**Two ω**	**Two Ratio Fixed ω**
10..11	0.0069	0.110	0.011	0.0116
11..12	0.0044	0.110	0.011	0.0116
12..13	9.2158	0.110	0.011	0.0116
13..14	0.8129	0.110	0.956	1.0000
14..15	1.4522	0.110	0.956	1.0000
15..1	1.7828	0.110	0.956	1.0000
15..2	0.7334	0.110	0.956	1.0000
14..3	0.4714	0.110	0.956	0.0116
16..5	0.0904	0.110	0.011	0.0116
12..17	3.8865	0.110	0.0116	0.0116
17..6	0.0001	0.1100	0.0116	0.0116
17..7	0.0069	0.110	0.011	0.0116
11..8	0.0045	0.110	0.011	0.0116
	3.2360	0.110	0.011	0.0116
P *	33	18	19	18
L**	-3685.47113	-3685.47	-3688.72	-3689.15

*P value denotes the number of parameters used for each model.
**L is equal to the log likelihood value of each model.

APPENDIX D

QUANTITATIVE RT-PCR Ct VALUES FOR EVERY GENE AND REPLICATES

Table: D. 1. Quantitative RT-PCR Ct values for every gene and replicate are given below.

	Rcd-1 r Region	*CG13102*	*Gapdh2*
Curly replicate1	31.48	15.50	20.81
Curly replicate 2	25.72	19.32	27.27
Curly replicate 3	17.55	20.02	17.37
Noncurly replicate1	21.79	19.32	16.82
Noncurly replicate2	16.66	18.25	19.92
Noncurly replicate3	18.89	19.31	18.06

APPENDIX E

ALIGNMENT OF Rcd-1 CARBOXY TERMINUS WITH HUMAN Rcd-1 REGION USED TO RAISE ANTIBODIES

```
Rcd-1human        ICQTYERFSHVAMILGKMVLQLSKEPSARLLKHVVRCYLRLSDNPRAREALRQCLPDQLK 60
Rcd-1parental     ICQTYERFSHVAITLGKMVIQLAKDPCARLLKHVVRCYLRLSDNTRARKALGQCLPDQLR 60
Rcd-1retro        ICENHDRFSQVAITLGKMVIHMLKFPCLRVLKHVVRCYLLLTENARARSALRVCLPDLLR 60
                  **:.::***:**: *****::: * *. *:********* *::*.***.**  **** *:
Rcd-1human        DTTFAQVLKDDTTTKRWLAQLVKNLQEGQVT-------DPRGIPLPP---- 100
Rcd-1parental     DGTFALCLQEDKSTKQWLQMLLKNLELGAT---------PQQIGMSPLGS- 101
Rcd-1retro        DGTFTSLVQHDTCTKQWLQMLLKNLQTNAV--------NPMGSS------- 96
                  * **:  ::.*. **:**  *:***: .
```

Figure: E.1 Alignment of Rcd-1 and Rcd-1r carboxy terminus with the human Rcd-1 region used to raise antibodies

.

REFERENCES

Arguello, J. R., Y. Chen, et al. (2006). "Origination of an X-linked testes chimeric gene by illegitimate recombination in Drosophila." PLoS Genet **2**(5): e77.

Ausubel, F. M., R. Brent, et al. (1988). Current protocols in molecular biology Canada, John Wiley & Sons.

Bai, Y., C. Casola, et al. (2007). "Comparative Genomics Reveals a Constant Rate of Origination and Convergent Acquisition of Functional Retrogenes in Drosophila." Genome Biology **8**(1): R11.

Betrán, E. and M. Long (2003). "*Dntf-2r:* a young Drosophila retroposed gene with specific male expression under positive Darwinian selection." Genetics **164**: 977-988.

Betrán, E., K. Thornton, et al. (2002). "Retroposed new genes out of the X in Drosophila." Genome Res. **12**(12): 1854-1859.

Blumer, N., K. Schreiter, et al. (2002). "A new translational repression element and unusual transcriptional control regulate expression of don juan during Drosophila spermatogenesis." Mech Dev **110**(1-2): 97-112.

Castrillon, D. H., P. Gonczy, et al. (1993). "Toward a molecular genetic analysis of spermatogenesis in Drosophila melanogaster: characterization of male-sterile mutants generated by single P element mutagenesis." Genetics **135**(2): 489-505.

Chen, X., M. Hiller, et al. (2005). "Tissue-specific TAFs counteract Polycomb to turn on terminal differentiation." Science **310**(5749): 869-72.

Clark, A. G., M. B. Eisen, et al. (2007). "Evolution of genes and genomes on the Drosophila phylogeny." Nature **450**(7167): 203-218.

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

Esnault, C., J. Maestre, et al. (2000). "Human LINE retrotransposons generate processed pseudogenes." Nat Genet **24**(4): 363-7.

Garces, R. G., W. Gillon, et al. (2007). "Atomic model of human Rcd-1 reveals an armadillo-like-repeat protein with in vitro nucleic acid binding properties." Protein Sci **16**(2): 176-88.

Hiller, M., X. Chen, et al. (2004). "Testis-specific TAF homologs collaborate to control a tissue-specific transcription program." Development **131**(21): 5297-308.

Hiller, M. A., T. Y. Lin, et al. (2001). "Developmental regulation of transcription by a tissue-specific TAF homolog." Genes Dev **15**(8): 1021-30.

Hiroi, N., T. Ito, et al. (2002). "Mammalian Rcd1 is a novel transcriptional cofactor that mediates retinoic acid-induced cell differentiation." EMBO J. **21**(19): 5235-5244.

Hwa, J. J., A. J. Zhu, et al. (2004). "Germ-line specific variants of components of the mitochondrial outer membrane import machinery in Drosophila." FEBS Lett **572**(1-3): 141-6.

Kalamegham, R., D. Sturgill, et al. (2007). "Drosophila mojoless, a retroposed GSK-3, has functionally diverged to acquire an essential role in male fertility." Mol Biol Evol **24**(3): 732-42.

Kempe, E., B. Muhs, et al. (1993). "Gene regulation in Drosophila spermatogenesis: analysis of protein binding at the translational control element TCE." Dev Genet **14**(6): 449-59.

Long, M. and C. H. Langley (1993). "Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila." Science **260**(5104): 91-5.

McDonald, J. H. and M. Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila." Nature **351**(6328): 652-4.

Proschel, M., Z. Zhang, et al. (2006). "Widespread adaptive evolution of Drosophila genes with sex-biased expression." Genetics **174**(2): 893-900.

Rozas, J., J. C. Sanchez-DelBarrio, et al. (2003). "DnaSP, DNA polymorphism analyses by the coalescent and other methods." Bioinformatics **19**(18): 2496-7.

Schafer, M., R. Kuhn, et al. (1990). "A conserved element in the leader mediates post-meiotic translation as well as cytoplasmic polyadenylation of a Drosophila spermatocyte mRNA." Embo J **9**(13): 4519-25.

Schafer, M., K. Nayernia, et al. (1995). "Translational control in spermatogenesis." Dev Biol **172**(2): 344-52.

Sturgill, D., Y. Zhang, et al. (2007). "Demasculinization of X chromosomes in the Drosophila genus." Nature **450**(7167): 238-41.

Timakov, B. and P. Zhang (2001). "The hsp60B gene of Drosophila melanogaster is essential for the spermatid individualization process." Cell Stress Chaperones **6**(1): 71-7.

Ting, C. T., S. C. Tsaur, et al. (2000). "The phylogeny of closely related species as revealed by the genealogy of a speciation gene, Odysseus." Proc Natl Acad Sci U S A **97**(10): 5313-6.

Tripoli, G., D. D'Elia, et al. (2005). "Comparison of the oxidative phosphorylation (OXPHOS) nuclear genes in the genomes of Drosophila melanogaster, Drosophila pseudoobscura and Anopheles gambiae." Genome Biol **6**(2): R11.

Vinckenbosch, N., I. Dupanloup, et al. (2006). "Evolutionary fate of retroposed gene copies in the human genome." Proc Natl Acad Sci U S A **103**(9): 3220-5.

Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Comput. Appl. Biosci. **13**(5): 555-556.

Yuan, X., M. Miller, et al. (1996). "Duplicated proteasome subunit genes in Drosophila melanogaster encoding testes-specific isoforms." Genetics **144**(1): 147-57.

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

Taniya Muliyil had a passion for biology which incited her to earn her Undergraduate and Graduate degree in Microbiology and Biotechnology respectively from India. Later, she joined the University of Texas at Arlington in August 2005 and completed her degree Master of Science in Biology in December 2007. In future she is determined to pursue her career in research.

.