SEQUENCE EVOLUTION OF RECURRENTLY RECRUITED RETROPOSED GENES IN *DROSOPHILA* AND THEIR POSSIBLE ROLE IN MEIOTIC DRIVE

by

CHARLES DAVID TRACY

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

May 2009

ACKNOWLEDGEMENTS

I would like to thank my supervising professor Dr. Esther Betrán for her guidance and for constantly pushing me to better myself during the course of my studies as a masters candidate. I would also like to thank Dr. Cédric Feschotte and Dr. Pawel Michalak for taking the time and effort to serve on my supervising committee and for their invaluable comments on my research.

I am thankful for all the teachers I have had as a student whom are too numerous to name. I feel that, in every class, I have learnt valuable things that have change the way I think and contributed to make me who I am.

Special thanks are in order for the other members of the Betrán lab especially Javier, Mansi, Mehran, Claudio and Taniya, for their support and advice during my time working in the lab.

Also I want to thank my office-mates John Morse, Robert Makowsky, and Christian Cox along with office neighbor Brian Fontenot who are always good for a lively discussion/argument and even better for a hearty laugh.

Finally, I am grateful for my family who has always been there for me. I love you all.

April 14, 2009

ABSTRACT

SEQUENCE EVOLUTION OF RECURRENTLY RECRUITED RETROPOSED GENES IN *DROSOPHILA* AND THEIR POSSIBLE ROLE IN MEIOTIC DRIVE

Charles Tracy, M.S.

The University of Texas at Arlington, 2009

Supervising Professor: Esther Betrán

Gene duplications are a valuable source of genetic information that can evolve under positive selection creating a new gene function without affecting the original function.

A gene duplication mechanism is retroposition. Retroposed copies of genes (retrogenes) are created by reverse transcription of a mRNA into the host organism's genome producing a new sequence that has the same protein coding capacity as the parental gene but lacks introns and regulatory regions. *Ran* and *Dntf-2* are genes involved in nuclear transport that have given rise to retrogenes three times in the *Drosophila* genus.

Recently, genes involved in nuclear transport such as *Ran* and *Dntf-2* were implicated in playing a major role in a chromosomal segregation distortion system in *D. melanogaster*. This thesis provides evidence of positive selection acting on the retrogenes and discusses the potential role of retroposed nuclear transport genes in segregation distortion.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Gene Duplication

Gene duplications have occurred throughout the evolutionary history of all organisms and provide genomes with a source of genetic information that over time can be changed through the process of fixation of random mutations. Mutations that become fixed by natural selection over a period of time might create a gene with a novel function (Graur and Li, 2000). Gene duplications can occur by wide range of mechanisms including but not limited to unequal crossover, or whole genome duplication followed by subsequent loss of neutral or deleterious genetic material, and retroposition (Lynch and Conery, 2000; Long et al., 2003). Regardless of the mechanism involved, a duplicated gene is initially often redundant and potentially expendable, a situation frequently leading to nonfunctionalization and likely followed by expulsion of the pseudogene from the genome (Lynch and Conery, 2000). For a new gene to become indispensable and be maintained by natural selection it must undergo either subfunctionalization or neofunctionalization process wherein it can either retain some of the functions of the parental gene or evolve a new function altogether (Lynch and Conery, 2000; Prince and Pickett, 2002). For a new gene to find a new genomic niche it will have to experience a period of relaxation of selection and/or positive selection wherein DNA substitutions alter the gene's coding region leading to a novel amino acid sequence that could confer increased fitness to the host organism if a suitable expression pattern is present (Jones and Begun, 2005). Once a new beneficial gene is created and expressed, it can be maintained by purifying selection (Long et al. 2003; Jones and Begun, 2005). Because duplicate genes have played a major role in the evolution of all organismal genomes since the genesis of life, it

is of great biological significance to locate duplicates and study how these genes have evolved essential roles within the organisms they inhabit. The remainder of this chapter will focus on the mechanism and hallmarks of gene retroposition, a discussion of retroposed genes in *Drosophila*, and finally an introduction to two nuclear transport genes that have given rise to multiple retrogenes in the *Drosophila* phylogeny. Chapters two and three report on research involving the evolution of functional retrogenes stemming from the aforementioned nuclear transport genes. Finally, chapter four discusses the hypothesized role that duplicated nuclear transport genes might have in genetic conflicts, specifically related to what is known about the segregation distortion (SD) system in *D. melanogaster*.

1.2 Retrogenes

Retrogenes arise when messenger RNAs are reverse transcribed into DNA by a reverse transcriptase/endonuclease (RT/EN) enzyme encoded by non-LTR retrotransposons and are randomly inserted into the genome (Brosius, 1991; Feng et al., 1996). Reverse transcription begins when a cytosolic RT/EN captures a processed mRNA and escorts the mRNA to the cell nucleus where the genomic DNA resides. Once in the genome, the endonuclease domain of the RT/EN enzyme randomly nicks one strand of the genomic DNA leaving an exposed 3' hydroxyl residue. A DNA strand complementary to the mRNA is synthesized by the reverse transcriptase domain of the RT/EN enzyme beginning at the exposed 3' OH group (Feng et al. 1996). It remains unclear how the second DNA strand is synthesized. Retrogenes have some hallmark attributes including the absence of introns and a poly A tail just as the template mRNA. Another common attribute is a direct flanking repeat at both ends of the gene because of the staggered cut made by the RT/EN enzyme (Brosius 1991). The 3' poly A tail and direct flanking repeats are only intact and visible for very young retrogenes as there is no selective pressure for their maintenance.

Most retrogenes are believed to become pseudogenes mainly because they are supposed not to carry regulatory regions with them (Bai et al. 2007). For a gene to be seen by natural selection it must be expressed at some stage or in some tissue in the organism where it resides. Retrogenes are supposed to be rarely, if ever, reverse transcribed with an intact promoter to guide expression because few genes have internal promoters. In Drosophila there is no evidence that retrogenes carry over upstream regulatory regions (Bai et al. 2007). However, it has recently been proposed that this might be occurring often in mammals (Okamura and Nakai, 2008). In fact one of the classic mammalian retrogenes, *Pgk2* is believe to have carried over general promoter from the parental gene and have evolved more recently the testis-specific expressions (McCarrey, 1994).

About the rate at which retrogenes form, Bai et al. (2007) found evidence of 94 functional retrogenes in *D. melanogaster* whole genome sequence data and doing comparative work using the eleven additionally sequenced *Drosophila* species estimated that functional retrogenes originated and became fixed in the genome at a fairly constant rate of 0.5 genes per million years per lineage. Their research showed that retroposition of genes is a common mechanism for origination of new genes in *Drosophila*.

Additional work has shown that retrogenes can evolve under positive selection to evolve novel functions. In 1993, Long and Langley demonstrated that a gene retroposition event could create an entirely new gene when inserted into an existing gene when they discovered *jingwei*. This gene was created when and *Adh* mRNA was reverse transcribed and inserted into a duplication of the *yellow emperor* gene thereby creating a new chimeric gene name *jingwei* (Long and Langley, 1993; Wang et al. 2000). Furthermore, sequence analyses indicate *jingwei* has experienced accelerated evolution that was paralleled by other independently derived *Adh* chimeras (Long and Langley, 1993; Jones and Begun, 2005).

While chimeric gene fusion is a nearly (mutations are still needed to adjust the retrogene to the right frame, i.e. in the case of jingwei a splicing donor site in the right frame had

to be created) instant source of new novelties, other retroposition events that do not get inserted into another gene might need to evolve “de novo” transcription or insert close to a regulatory region. There are many of these examples in *Drosophila* (Bai et al. 2007). In this work, Bai et al. concluded that retrogenes' regulatory regions mostly do not represent a random set of existing regulatory regions. Selection in favor of retrogenes inserted in male testis neighborhoods and at the sequence level to produce testis expression is postulated to have occurred. Dorus et al. (2008) also showed an excess of sperm retrogenes inserted in testis neighborhoods.

Interestingly, many of the non-chimeric retrogenes discussed above are X chromosome to autosome duplications and are highly expressed in male germline tissues (Betran 2002; Bai et al. 2007). It seems that a significant excess of genes have been retroposed from the X chromosome to autosomal locations, and acquired male germline expression patterns. This pattern is different from the pattern observed for retrotransposable elements despite the fact that the retroelement machinery is used to produce retrogenes. This reveals that is likely selection that underlies the patter in *Drosophila* (Bai et al. 2007). This pattern has contributed to the demasculinization of the X chromosome observed genome wide (Parisi et al.2003). Some of these genes have important fertility functions. Kalamegham et al. (2006) described the retrogene *mojoless* and showed that is required for male germ line survival. Yuan et al. (1996) characterized an X to autosome retroposed gene that encodes a testes-specific proteasome component dubbed *Pros28.1A*. Other autosomal retrogenes have been proven to encode novel sperm components in *Drosophila*. Dorus and colleagues (2008) reported four proteins of recent retrotranspositions (two X to autosome) present in sperm, with three of the four having enriched testes expression relative to the parental gene.

X to autosome retrogenes have been discovered not only in *Drosophila* but also in mammals. Research done by Emerson et al. (2004) indicate that a disproportionately high number (299% excess in humans and 309% excess in mice) of functional retroposed genes have been recruited to autosomes from the X presumably to carry out male specific functions.

This pattern was not found for retropseudogenes revealing that selection and not mutation is the underlying force that explains this excess. Again they often have spermatogenesis functions. *Utp14c* and *Utp14b* are X to autosome duplicates required for spermatogenesis and fertility in humans and mice respectively (Rohozinski et al. 2006; Bradley et al. 2004).

Because so many examples of X to autosome retropositions have been documented, with many acquiring testes biased expression and this can not be explain by mutational biases (see above), there appears to be some selective pressures involved in recruitment of male specific genes out of the X chromosome. But what kink of selective pressures would cause such an abundance of X chromosome to autosome gene duplications?

The two dominant hypotheses are not mutually exclusive and point to enhanced fitness of males that have exported genes necessary for male meiosis to autosomes. The first hypothesis was put forth by McCarrey for *Pgk2* (1994) and states that because the X chromosome is inactivated by *XIST* transcripts during male meiosis (Richler et al. 1992), it becomes advantageous for genes involved in male meiosis to be in autosomes where they can be expressed during meiotic divisions. The second hypothesis points to a sexual antagonism model wherein for dominant mutations that are beneficial to males and have deleterious effects on females would be more likely to be found in autosomes because the X spends two-thirds of its time in females (Wu and Xu, 2003). However, the way the retrogenes studied in this work are evolving, does not fit either hypothesis well. The acquisition of a completely novel function (possibly meiotic drive function; see below) that is under recurrent positive selection fits better the observed data.

1.3 *Ran* and *Dntf-2* Derived Retrogenes

Recently Bai et al. (2007) revealed convergent duplications of two genes involved in nuclear transport in different lineages of Drosophila. *Dntf-2* and *Ran* seem to have given rise to retroposed copies (i.e. retrogenes; (Brosius 1991) three independent times. *Dntf-2* gave rise to

a retrogene (*Dntf-2r*) that is present in the 2L chromosomal arm of 4 species of the *D. melanogaster* complex (*D. melanogaster, D. simulans, D. mauritiana* and *D sechellia*; (Betrán and Long 2003), a retrogene in the *D. ananassae* lineage (located in the arm that corresponds to 3L arm in *D. melanogaster*) and another independently originated retrogene in the lineage leading to *D. grimshawi* (located in the arm that corresponds to 3L arm in *D. melanogaster*). *Ran* seems to have given rise to retrogenes three times in the same lineages as *Dntf-2* as shown in figure 1.1. It gave rise to *Ran-like* that is present in all the species of the *D. melanogaster* subgroup and located in the 3L arm. *D. ananassae* and *D. grimshawi* also have a different *Ran* retrogene. These retrogenes are located in the arm that corresponds to 2L and 3L arm in *D. melanogaster* respectively.

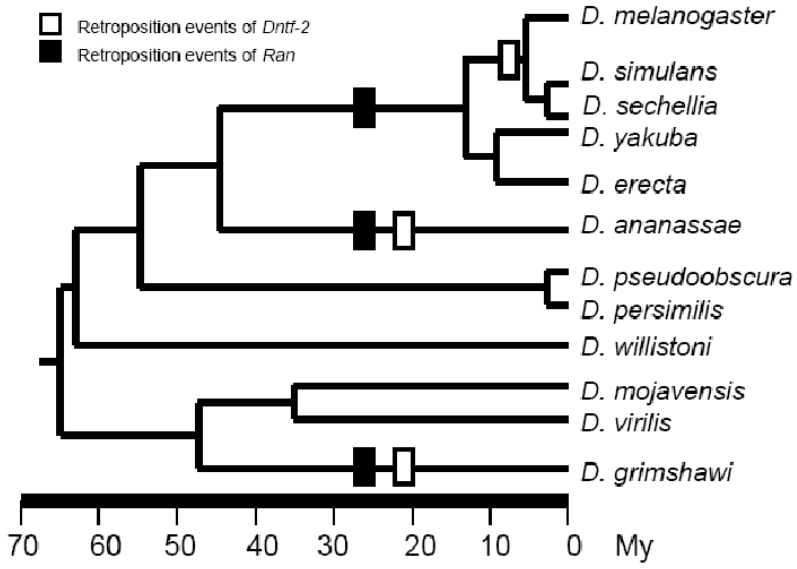

Figure 1.1 *Drosophila* phylogeny indicating retroposition events of *Dntf-2* and *Ran*. (Figure modified from Bai et al. 2007)

The acquisition of *Dntf-2* and *Ran* duplicates by retroposition in some lineages may have an adaptive meaning, particularly if they overlap in their expression. All six independent retroposition events occurred from an X to autosome locations that have been claimed to be favored by selection. Positive selection might occur whenever the gene recruits a male germline expression due to either male germline X inactivation or sexual antagonism as described in section 1.2. Alternatively, the convergent recruitment of said retrogenes could be the result of genetic conflict to be described in section 1.5 and discussed in Chapter 4.

1.4 *Ran*, *Ntf-2*, and Nuclear Transport

Ran and Ntf-2 proteins physically interact and play a central role in transport of proteins to the nucleus (Ribbeck et al. 1998). Ran is a member of Ras superfamily. It exists in GDP bound inactive form and GTP bound active form. RanGDP predominantly localizes in the cytoplasm and RanGTP in the nucleus. The resulting cytoplasm-nuclear gradient of RanGDP-RanGTP is important for the import and export of cargo proteins across the nuclear membrane. Ntf-2 is a transport protein that interacts with RanGDP in the cytoplasm and carries it across the nuclear pore into the nucleus (Quimby et al. 2000). Once in the nucleus, a catalytic enzyme RanGEF (Ran GTPase Exchange Factor), also called RCC1, converts the Ran from the GDP bound to the GTP bound form. RanGTP binds to importin β and induces conformational changes that lead to the dissociation of importin α/β heterodimer and release of the cargo protein. RanGTP ensures the release of cargo proteins in precise spatial and temporal pattern for the proper orchestration of downstream functions. RanGTP bound to importin β is transported out of the nucleus (Isgro and Schulten 2007). RanGTP is also needed for assembly of export complexes and it is transported with these complexes to the cytoplasm (Kusano et al. 2003; Matsuura and Stewart 2004). Once in the cytoplasm, another catalytic enzyme - RanGAP (Ran GTPase Activating Protein) hydrolyses RanGTP into RanGDP (Kusano et al. 2002). See figure 1.2 for a summary of nuclear transport.

In addition to the transport functions, RanGTP concentration gradients are required during normal cellular divisions. Multiple experiments have indicated a cloud of RanGTP is generated around chromosomes in vertebrate somatic cells as well as in *X. laevis* egg extracts. This gradient is required for spatial organization of the spindle apparatus and targeting microtubules toward kinetochores. Furthermore, the Ran GTP gradient is required for nuclear envelope assembly following cell division and also has a role in nuclear envelope assembly around sperm chromatin in *X. laevis* egg extracts (Clarke and Zhang, 2008).

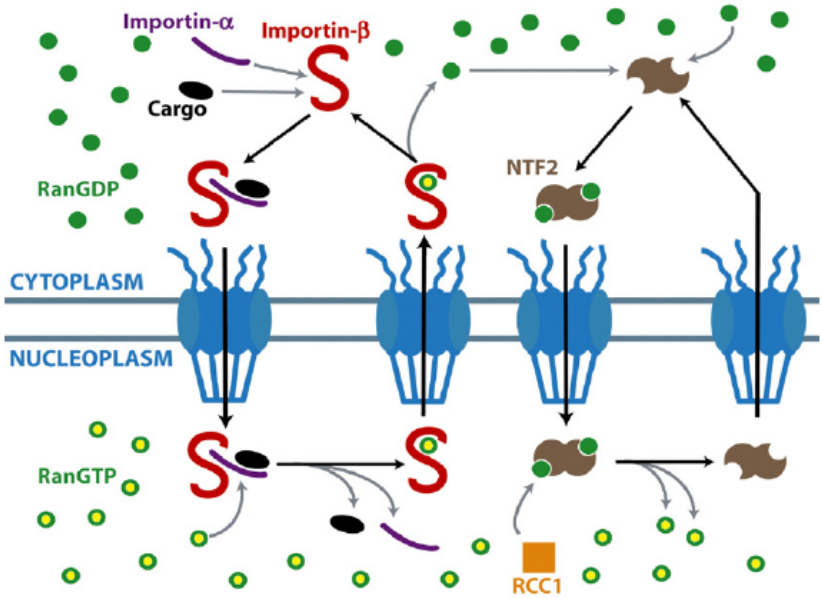

Figure 1.2 Nuclear transport schema (Figure modified from Isgro and Schulten, 2007)

1.5 Genetic Conflict – Meiotic Drive

The *D. melanogaster Dntf-2r* and *Ran-like* exhibit a male germline biased expression (Betrán and Long 2003; Chintapalli et al. 2007). Thus it is possible that the multiple parallel retropositions to autosomes of *Dntf-2* and *Ran* have been driven by a selective advantage for

the males that had both additional copies because of the need to retain nuclear transport function during spermatogenesis and X inactivation. It could also be explained by sexual antagonism as outlined above. Alternatively, the multiple duplication events of nuclear transport genes *Dntf-2* and *Ran* could be explained by their role in meiotic drive. Meiotic drive systems are selfish systems that increase their frequency at which they are transmitted to the next generation. Usually, this is viewed as a genomic conflict because the driving chromosome is not necessarily the chromosome with the best alleles for the individual. Interestingly, there is direct evidence that genes involved in nuclear transport play a role in the *SD* segregation distortion system in spermatogenesis of *D. melanogaster*. In this system the *SD* chromosome is a second chromosome that is transmitted to the progeny between 95-100% of the time. It carries a main distorter locus, an insensitive responder locus and other loci that act as enhancers or modifiers of the drive. The main distorter, *Sd*, (see figure 1.3) is a truncated form of the nuclear transport gene *RanGAP* that (mis)localizes in nucleus (Kusano et al. 2003). It has been proposed that the truncated form of *RanGAP* (*Sd-RanGAP*), which retains catalytic activity, acts by hydrolyzing RanGTP into RanGDP in the nucleus, thereby disturbing the RanGDP/RanGTP gradient (Kusano et al. 2002). This perturbation disables development of sperm carrying the second chromosome with the responder sensitive allele. It has been described that the simple over expression of *RanGAP* causes segregation distortion. Over expression of other nuclear transport genes (i.e. *Ran* and *RanGEF*) has been shown to compensate for the distortion (Kusano et al. 2002).

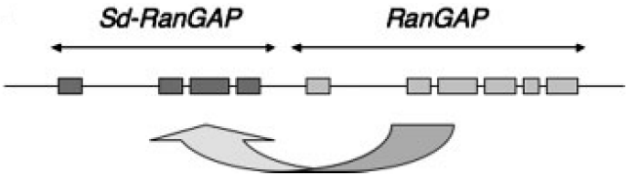

Figure 1.3 *Sd* locus in *D. melanogaster* (figure from Presgraves 2007)

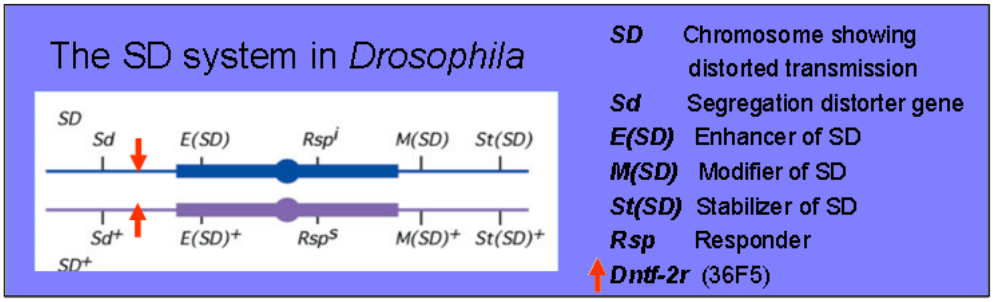

Figure 1.4 The *SD* system in *D. melanogaster* (figure modified from Presgraves 2007)

It has recently been suggested that many nuclear transport genes or their duplicates (i.e. *RanGAP*, nucleoporins and *Dntf-2r*) might also be involved in segregation distortion in Drosophila (Presgraves 2007). *Dntf-2r* is located within the *SD* region (36F cytological position in the left arm of the second chromosome in *D. melanogaster*) and linkage between the elements that play a role in distortion has been recognized to be an important feature of these systems (Kusano et al. 2003; Burt and Trivers 2006). The arms race between distorters and compensators could lead to the fixation of duplicate genes with male germline expression and fast gene evolution (Kusano et al. 2002; Burt and Trivers 2006; Presgraves 2007). All of these genes (*RanGAP*, nucleoporins and *Dntf-2r*) are evolving under positive selection (Betrán and Long 2003; Presgraves 2007; Presgraves and Stephan 2007). The selective advantage for the fixation, fast evolution and longevity of the duplicates will persist as long as this or other conflicts of similar nature remain.

The data presented below about the way *Dntf-2* and *Ran* retrogenes are evolving seem to fit better the acquisition of a new function (possibly meiotic drive function) by these genes.

CHAPTER 2

EVOLUTION OF *Dntf-2* DERIVED RETROGENES

2.1 Introduction

The selective regime operating on gene sequence evolution can be studied by a variety of approaches. One of the most common methods involves calculation of nonsynonymous/synonymous rate ratio ($\omega = K_A/K_S = d_N/d_S$) between orthologous sequences. K_A is defined as the proportion of nonsynonymous mutations per nonsynonymous site. Likewise, K_S is the proportion of synonymous mutations per synonymous site. Because mutations at synonymous sites are invisible to natural selection (i.e. do not lead to change in amino acid), these changes are allowed to accrue at a constant rate without changing the protein composition or decreasing the fitness of the host. In contrast, nonsynonymous mutations alter the amino acid sequence of the protein encoded by a gene and are under strong selective pressures. Comparison of the two rates (K_A and K_S) provides a powerful tool for understanding the molecular evolution of a gene wherein $\omega = 1$ is consistent with neutral evolution because synonymous and nonsynonymous mutation occur at the same nonselective rate. A ω ratio significantly less than 1 is indicative of purifying selection as a result of few amino acid altering substitutions in the coding sequence. Positive selection or adaptive evolution is inferred when ω is significantly greater than 1. This occurs when significantly more nonsynonymous substitutions have been fixed relative to synonymous changes in the same sequence and the occurrence of recurrent positive selected replacement substitutions is inferred (Goldman and Yang, 1994; Yang, 1998; Yang et al., 2000; Yang and Neilsen, 2000).

PAML (Phylogenetic Analysis by Maximum Likelihood) software developed by Ziheng Yang (1997) can be used to determine ω ratios for multiple branches in a phylogeny.

Hypothesis based a priory models can be compared in order to understand the selective pressures acting on different branches of a gene tree or multiple genes tree (Yang, 1998). If some lineages are fast evolving, site-specific models can be fitted to test whether some amino acids in the gene-based phylogeny are under positive selection. Specific models that test our a priory hypotheses will be discussed in greater detail in the following methods and results sections.

Similarly to PAML, HyPhy software (http://www.datamonkey.org) is capable of calculating ω ratios for multiple branches on a gene tree and for aligned codons in an effort to understand the selective pressures on genes in a given phylogeny. Random effects likelihood (REL) and fixed likelihood effects (FEL) are employed to detect sites under positive and negative selection without estimating parameters such as branch length (number of substitutions) or specifying site classes based upon ω rates a priori. Furthermore, the REL and FEL models allow synonymous and nonsynonymous rates to vary and synonymous and nonsynonymous rates to vary on a site to site basis (Kosakovsky Pond and Frost 2005). This approach makes the models more realistic. Specific analyses will be discussed in the following sections.

PAML and HyPhy softwares are powerful tools used in understanding how genes have evolved among many different species (i.e. when studying divergence data). The protein coding sequences used in these analyses are downloaded from FlyBase and represent the sequence of a gene in the sequenced isoline of a given species (Clark et al. 2007). However, polymorphism data can also be use as a different approach to detect the selection regime the gene is undergoing. Sequencing alleles of a gene from naturally occurring populations of different species and comparing synonymous versus nonsynonymous fixations and polymorphisms gives a more refined statistical method for determining selective pressures associated with a particular gene. McDonald and Kreitman (1991) conceived a test based on sequenced data from natural populations to test the neutral theory. Tajima (1989) proposed a

statistic, Tajima's D, to test the neutral theory for any gene in a single population by subtracting two different estimates of theta ($\theta=4N\mu$); one calculated from the number of polymorphic sites and the other the average nucleotide diversity for the gene in question. A value of zero is consistent with neutral evolution while positive and negative values might be indicative of balancing and either purifying or positive selection respectively (Tajima, 1989). A recent selective sweep would produce a negative Tajima's D value (Fay and Wu 2000). Population expansion or population structure can also explain these departures and this should be kept in mind. While Tajima's D uses sequence data from one species, the McDonald-Kreitman test uses sequence data from two or more species. In the McDonald-Kreitman test, if the gene in question is evolving under neutrality the ratio of nonsynonymous/synonymous replacements between species should equal the nonsynonymous/synonymous polymorphisms within species. An excess of replacement substitutions is consistent with recurrent adaptive evolution by positive selection (McDonald and Kreitman, 1991).

As introduced above, *Dntf-2* gave rise to a retrogene (*Dntf-2r*) that is present in the 2L chromosomal arm of 4 species of the *D. melanogaster* complex (*D. melanogaster, D. simulans, D. mauritiana* and *D sechellia*; (Betrán and Long 2003), a retrogene in the *D. ananassae* lineage (located in the arm that corresponds to 3L arm in *D. melanogaster*) and another independently originated retrogene in the lineage leading to *D. grimshawi* (located in the arm that corresponds to 3L arm in *D. melanogaster*).

Previous work on *Dntf-2r* by Betrán and Long (2003) has revealed the selective pressures acting on this retrogene in the *D. melanogaster* complex. A McDonald-Kreitman test using *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana* sequences indicated positive selection in the *Dntf-2r* lineages. PAML analyses and Tajima's D were also implemented but failed to show adaptive changes occurring in the *Dntf-2r* lineages. The gene is however, at the same time under purifying selection (Betrán and Long 2003).

Given the advent of whole genome sequencing and the completion of twelve *Drosophila* genomes, more research can be done to determine the mode of evolution of *Dntf-2* retroposed sequences. In the remainder of this chapter, PAML analysis of *Dntf-2* derived gene duplicates in the *D. melanogaster* complex *(Dntf-2r*), *D. ananassae* lineage (*Da_Ntf-2r*), and the *D. grimshawi* lineage (*Dg_Ntf-2r*) along with all *Drosophila Dntf-2* lineages will be discussed. In addition to the PAML analysis, Tajima's D and a McDonald-Kreitman test will be performed to understand how the *Dntf-2* derived retrogenes in the *D. ananassae* lineage have evolved.

2.2 Methods

2.2.1. PAML and HyPhy software analyses

The nucleotide sequences of the *Dntf-2* gene in the 12 sequenced Drosophila species were retrieved from FlyBase (Clark et al., 2007) along with the sequences of the ortholog *Dntf-2r* in *D. melanogaster*, *D. simulans*, and *D. sechellia*. These 15 sequences were aligned with each other and *Dntf-2* retroposed sequences from *D. ananassae* and *D. grimshawi* with Clustal W software (Thompson et al., 1994). Sequences were analyzed using the CODEML software package implemented in PAML 3.15 (Yang 1997). The tree provided was (((((((*Dntf-2r_D.simulans*, *Dntf-2r_D.sechellia*), *Dntf-2r_D.melanogaster*), ((*Dntf-2_D.simulans*, *Dntf-2_D.sechellia*), *Dntf-2_D.melanogaster*)), (*Dntf-2_D.yakuba*, *Dntf-2_D.erecta*)), (*Da_Ntf-2r_D.ananassae*, *Dntf-2_D.ananassae*)), (*Dntf-2_D.psedoobscura*, *Dntf-2_D.persimilis*)), *Dntf-2_D.willistoni*), ((*Dntf-2_D.virilis*, *Dntf-2_D.mojavensis*), (*Dg_Ntf-2r_D.grimshawi*, *Dntf-2_D.grimshawi*))) and the topology is seen in figure 2.1. Branch models were employed to determine selective pressures on the parental and retroposed sequences by calculation of ω ratios for each branch (Yang 1998). The free model allows every branch to evolve at a different rate, and the one rate model sets all branches to evolve at the same rate. Other models allow for differing ω rates for all retroposed sequences and *Dntf-2* branches (two ratio model), and independent ω ratios for each retroposed sequence along with one ratio for the *Dntf-2* branches

(four ratio model). A five ratio model was fit to test whether the branch following duplication of *Dntf-2r* in the *D. melanogaster* complex (Node 5…Node 6, (see figure 2.1)) experienced accelerated evolution relative to subsequent branches. A seven ratio model was created to test if *Dntf-2* branches that have given rise to a retrogene (*D. melanogaster* complex, *D. ananassae*, and *D. grimshawi*) have evolved with differing ω ratios relative to *Dntf-2* lineages that have not given rise to a retrogene. To resolve whether *Dntf-2r* branches evolve at different rates an eight ratio model was created to assign independent ω ratios to each *Dntf-2r* branch in the phylogeny. Finally, multiple null models were fit to the data wherein ω is fixed at one (ω=1) in *Dntf-2* retroposed lineages and compared with the best alternative model to detect positive/purifying selection. These were all a priori hypotheses to be tested. 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.

Next, site models of CODEML software implemented in PAML were used to uncover the possibility of positive selection acting on a few sites. Site-specific likelihood model pairs M1 (nearly neutral) and M2 (positive selection) along with M7 and M8 were applied to the sequences with the appropriate tree topology (Nielsen and Yang 1998; Yang et al. 2000). Models M1 and M7 do not allow for sites under positive selection and are compared to models M2 and M8 respectively, both of which allow for sites under positive selection. All models allow variable selective pressures among sites but fix rates among branches in the phylogeny. Model M2 is an extension of M1 by incorporation of an additional rate class $\omega_2 > 1$ (estimated from data with proportion p_2) to the two rate classes ($\omega_0 << 1$ and $\omega_1 = 1$) and proportions (p_0 and p_1) present in M1. M7 assumes a beta distribution for ω between 0 and 1 over all sites while M8 adds an additional site class (ω≥1) with ω estimated from the data. A likelihood ratio test (LRT) was performed for both pairs by calculating two times the log likelihood values and comparing this value to a χ^2 distribution with 2 degrees of freedom. Posterior probabilities of codons under

positive selection are computed in models M2 and M8 using Bayes Empirical Bayes when the LRT was significant. The tree provided for the site analyses for *Dntf-2r* was ((*Dntf-2r_D.simulans, Dntf-2r_D.sechellia*), *Dntf-2r_D.melanogaster*).

The same seventeen *Dntf-2* and *Dntf-2* retrogene aligned sequences along with the tree topology used for PAML branch-specific analyses were uploaded to the HyPhy software package available at http://www.datamonkey.org. The GA branch model was run to uncover when selection occurred in the phylogeny. The models tested in the GA branch test are similar to the branch-site models and the branch-specific models employed in the PAML software package. As in the branch-site models, branch lengths and substitution rates are calculated by maximum likelihood for the phylogeny and held constant while site-to-site rate variation is calculated for codons in the sequence alignment. Finally, individual branches are automatically partitioned into different discrete classes with each class having independent ω ratios until a general optimized model is found. While the PAML branch-specific models bin branches specified by the user *a priori*, the GA branch models bin branches into an increasing number of ω ratio classes until the most probable model is located (Kosakovsky Pond and Frost, 2005).

Dntf-2r sequences and tree topology ((*Dntf-2r_D.simulans*, *Dntf-2r_D.sechellia*), *Dntf-2r_D.melanogaster*) were uploaded to the HyPhy software package available at http://www.datamonkey.org. REL analysis was performed in an attempt to detect positively selected codons in the *Dntf-2r* phylogeny at a Bayes factor threshold of 50 which corresponds to a small P value (Kosakovsky Pond and Frost, 2005). P value is roughly equivalent to 1/Bayes factor (Kosakovsky Pond and Frost, 2005). FEL was also run to detect positively selected codons using a significance level of $P < 0.1$.

2.2.2. Strains and sequencing

Genomic DNA extractions were performed on single fly using Puregene kit. *Dntf-2* retroposed sequences were PCR amplified from genomic DNA from ten *D. ananassae* strains (14024-0371.16, 14024-0371.17, 14024-0371.18, 14024-0371.25, 14024-0371.30, 14024-

0371.31, 14024-0371.32, 14024-0371.32, 14024-0371.33, 14024-0371.34, and 14024-0371.35), and four *D. atripex* strains (14024-0361.00, 14024-0361.01, 14024-0361.02, and 14024-0361.03). These strains were acquired from UC San Diego *Drosophila* stock center. Oligoprimers 5'- ATG CCT CTC AAT CCC CAC -3' and 5'- TTA TTC CGT GTC GTG GAT ATT C -3' were used for amplification of the retroposed *Da_Ntf-2* gene in all species. PCR products were sequenced from both strands using an automated DNA sequencer using fluorescent HiDi terminators. PCR products of individuals heterozygous for this gene were cloned and a clone was sequenced to establish the haplotypes. McDonald-Kreitman test (McDonald and Kreitman 1991) was performed using the polymorphism data obtained for the retroposed *Dntf-2* sequence in *D. ananassae*, and *D. atripex*. Sequenced products were aligned using Clustal W (Thompson et al, 1994) and imported into DnaSP 4.0 (Rozas et al. 2003) to perform the McDonald-Kreitman test and for calculation of Tajima's D statistic.

2.3 Results

2.3.1. PAML and HyPhy software analyses

To understand the mode of evolution of *Dntf-2* and its retrogenes we performed sequence analyses using PAML software as described in the Materials and Methods. Results of the branch-specific PAML analyses for *Dntf-2* genes and *Dntf-2* retrogenes appear in table 2.1 with figure 2.1 to be used as a reference for branch specifications. The free-ratio model (data not shown, log-likelihood (ℓ = -2507.6789), and parameters (p = 65) was found to be significantly better (P = 1.787×10^{-13}) than the one ratio model indicating that Ka/Ks ratios vary for the different lineages. The two-ratio model that estimates one ω ratio for *Dntf-2* branches and one ratio for all *Dntf-2* retrogene branches was found to fit the data better than the one-ratio model (P<1.110×10^{-16}) indicating differing ω rates for retroposed sequences relative to non-retroposed sequences. A four-ratio model was implemented to allow differing rates of evolution for the three retrogenes lineages (*Dntf-2r* clade, *D. ananassae Ntf-2r*, and *D. grimshawi Ntf-2r*) and *Dntf-2*. This model is significantly better than the two-ratio model (P = 8.390×10^{-5}) indicating

different rates of evolution among the recurrently recruited retrogenes. The retrogene in the *D. melanogaster* complex is evolving the fastest. All the retrogene rates are much higher (3 to 21 times higher) than the parental rate. This can be explained by positive selection acting on the retrogenes or relaxation of constraint in the retrogene lineages. We have evidence of positive selection acting on *Dntf-2r* provided by the McDonald-Kreitman test (Betrán and Long 2003) but no evidence yet for other lineages. However, ω ratio is estimated in *D. ananassae* lineage to be very fast (i.e. 0.3309). McDonald-Kreitman test was performed for this gene in *D. ananassae* and close related species (see below). In addition, we fit a five-ratio model to the data to uncover the selective pressures acting on the newly acquired retrogene following duplication in *D. melanogaster* complex. This model did not produce a significantly better fit when compared to the simpler four-ratio model (P = 0.4319). The seven- and eight-ratio models also failed to statistically improve the four-ratio model indicating that evolutionary pressures on the original *Dntf-2* sequences did not change after duplication and that *Dntf-2r* sequences have been under similar selective pressures. Four models were constructed with ω fix to one to determine if the estimated ω ratio for every gene was significantly less than one (i.e. purifying selection is acting). All four models were significantly worse at fitting the data than the four ratio model revealing purifying selection in all the lineages. These comparisons are shown in Table 2.1 except for *Dntf-2* lineages and *D. grimshawii* retrogene lineage (data not shown). This reveals that all ω values in the phylogeny are significantly less than one revealing purifying selection in all lineages.

Table 2.1 Results of PAML *Dtf-2r* branch-specific analyses

Model	l	*p*	$\hat{w}_{Dntf\text{-}2}$	$\hat{w}_{ms}$	$\hat{w}_{dup}$	$\hat{w}_{ana}$	$\hat{w}_{grim}$
One ratio	-2570.869	34	0.0533	0.0533	0.0533	0.0533	0.0533
Two ratio	-2531.728	35	0.0243	0.2657	0.2657	0.2657	0.2657
Four ratio	-2522.342	37	0.0247	0.5311	0.5311	0.3309	0.0754
Five ratio	-2522.033	38	0.0247	0.6235	0.3655	0.3310	0.0754
Four ratio null	-2524.291	36	0.0248	1	1	0.3328	0.0758

. Log-likelihood (ℓ) and number of parameters (p) along with ω estimates for branches are shown. Abbreviations ms, dup, ana and grim refer to *Dntf-2r* in the *melanogaster* subgroup, the branch following gene duplication (in *melanogaster* subgroup [node 5 to node 6 in figure 2.1]), *D. ananassae* retrogene, and *D. grimshawii* retrogene respectively.

Site models (pairs M1 and M2, M7 and M8) were also fitted to the data to test for positive selection acting on particular sites of *Dntf-2r* (See Materials and Methods). We have evidence of recurrent positive selection acting on *Dntf-2r* provided by the McDonald-Kreitman test (Betrán and Long 2003) and would like to see if we can decide at what sites that selection has been acting. From these results, see table 2.2, positive selection acting on any sites of the retroposed *Dntf-2* sequences could not be inferred because of insignificant likelihood ratio tests between the two pairs of models tested (M1 versus M2 $2\Delta\ell = 1.835$, with d.f. = 2, P = 0.3995; M7 versus M8 $2\Delta\ell = 1.848$, with d.f. = 2, P = 0.397).

Table 2.2 Results of PAML *Dntf-2r* site-model analyses.

Model	p	l	Parameter Estimates	Postively Selected Sites
M1: Neutral	7	-705.795	$p_0 = 0.42$ $p_1 = 0.58$	NA
M2: Selection	9	-704.878	$p_0 = 0.65$, $p_1 = 0$, **$p_2 = 0.35$, $w_2 = 2.41888$**	47, 68, 87, 116 (at 0.6 < P < 0.7)
M7: beta	7	-705.802	p = 0.00748, q = 0.005	NA
M8: beta + ω	9	-704.878	p = 0.005, q = 1.805, **$p_2 = 0.348$, $\omega_2 = 2.42$**	47, 68, 87, 116 (at 0.7 < P < 0.8)

Log-likelihood (l) and number of parameters (p) along with ω estimates and proportions (p) are shown.

The results from HyPhy software analyses are highly congruent with results from PAML analyses. Estimates of ω from GA-branch analysis, shown in figure 2.1, are very similar to

results from PAML branch-specific models especially for *Dntf-2r* branches in the *D. melanogaster* subgroup where both software analyses predict ω to be around 0.53. While this is not consistent with positive selection, this is several orders of magnitude larger than estimates for *Dntf-2* branches and indicative of relaxed constraint on *Dntf-2r*. Also, estimates for *Dntf-2* genes have very low ω estimates that are on par with PAML estimates for non-retroposed sequences, as would be expected from any highly conserved gene. FEL analysis was unable to detect positively selected codons at any significant level ($P < 0.1$). However, REL analysis detected two codons (47 [Bayes factor = 52.9636], 87 [Bayes factor = 52.6307]) that are likely under positive selection. It is interesting to note that these two codons were also uncovered as likely to be under positive selection by PAML site-specific models M2 and M8 but with much less confidence. These positively selected codons encode for amino acids that are identical in *Dntf-2* but are different for every *Dntf-2r* gene in the melanogaster subgroup (Figure 2.2). Codon 47 encodes for histidine in the parental *Dntf-2* genes and *Dntf-2r* in *D. sechellia* but has been replaced by a conservative arginine and a semi-conservative asparagine in *D. simulans* and *D. sechellia* respectively. Amino acid 87 has remained asparagines in *Dntf-2*, non-conservative changes to phenylalanine and isoleucine have occurred *D. sechellia* and *D. melanogaster* respectively. Though some changes are non-conservative amino acid replacements, they are not in regions of the protein that are known to interact with RanGDP indicating that these proteins are likely capable of interaction. Amino acids that interface with nucleoporin FxFG repeats (Stewart et al. 1998; Bayliss et al. 2002; Cushman et al. 2004) have experienced more numerous changes indicating decreased or altered interactions.

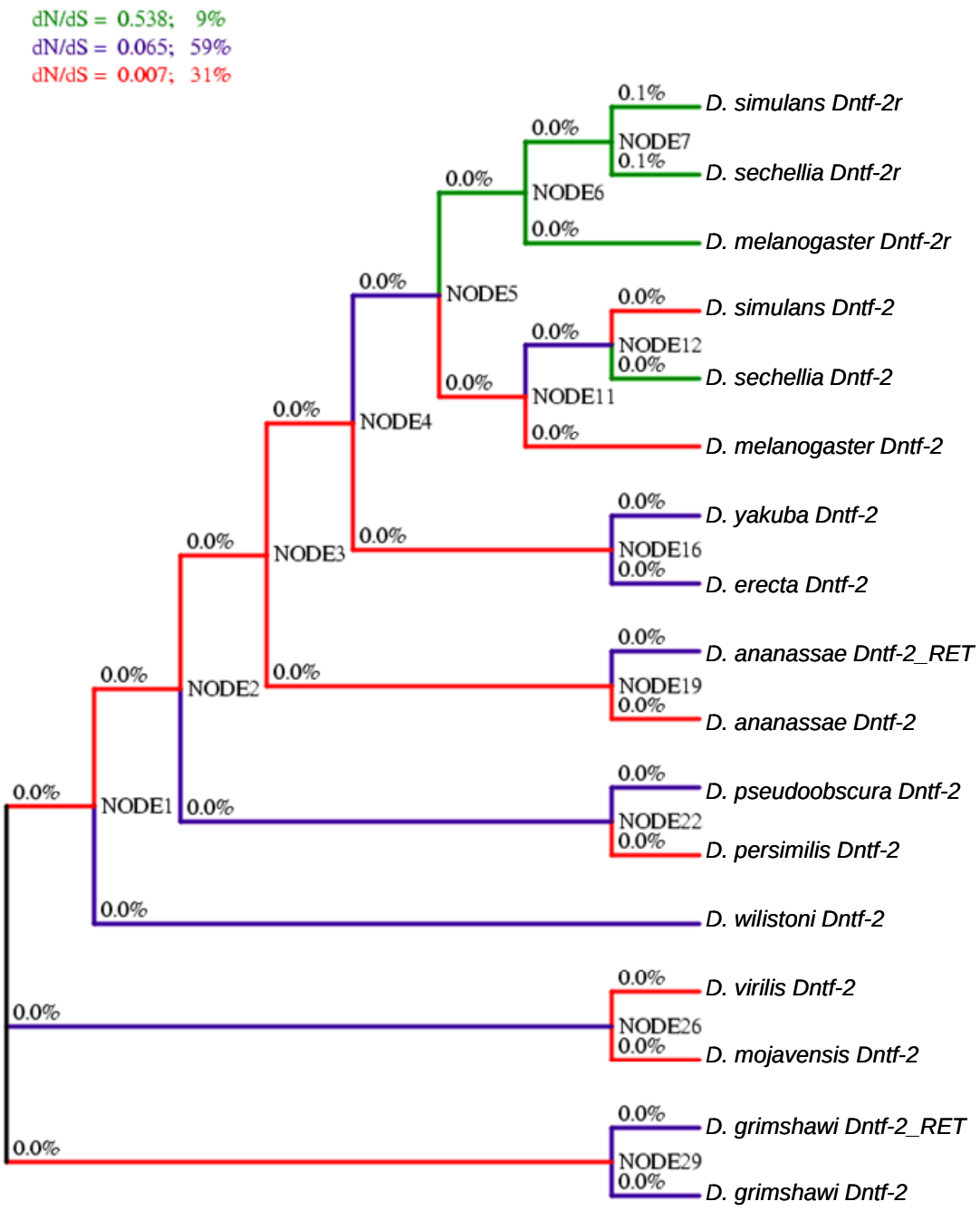

Figure 2.1 Results of GA-branch *Dntf-2r* analysis from HyPhy software package. RET refers to retroposed *Dntf-2r* sequences *Da_Dntf-2r* and *Dg_Dntf-2r* in *D.ananassae* and *D.grimshawi* respectively.

Figure 2.2 *Dntf-2* and *Dntf-2r* amino acid alignment. Red and green amino acids interact with RanGDP, Blue and green amino acids interact with importins. Black triangle points to positively selected site.

2.3.2. Polymorphism data

Polymorphism data for the retroposed *Ntf-2* gene in *D. ananassae* (n = 12) and *D. atripex* (n = 7) isofemale lineages was used to perform a McDonald-Kreitman test. The results reveal a statistically significant excess (two-tailed Fisher's exact test P = 0.001661) of replacement substitutions (Table 2.3) which is consistent with recurrent positive selection acting at the protein level.

Table 2.3 McDonald-Kreitman test for the retroposed *Dntf-2* gene in *D. ananassae* and *D. atripex*.

	Fixed	**Polymorphic**
Replacement	36	2
Synonymous	27	16

Two-tailed Fisher's exact test P = 0.001661**.

From the polymorphism data used for the McDonald-Kreitman test, one sequence per isoline was used to calculate Tajima's D statistic. Tajima's D statistic was calculated to be -0.89195 ($P > 0.1$) for the ten *D. ananassae* sequences. The negative value of Tajima's D statistic indicates an excess of low frequency polymorphisms relative to the average number of polymorphisms in isoline comparisons in the data set (Tajima, 1989). For the four *D. atripex* sequences, Tajima's D equals 0.67384 ($P > 0.1$). A positive value is generally indicative of population structure or is under balancing selection (Tajima, 1989). However, both tests were unable to reject the null hypothesis of neutrally evolving genes.

2.4 Discussion

An updated analysis of *Dntf-2* and retroposed sequences evolution using 12 sequenced Drosophila genomes has been performed. Ka/Ks analyses have revealed a 20x and 13x increase in the evolutionary rate of retroposed *Dntf-2* genes in the *D. melanogaster* subgroup and the *D. ananassae* lineage respectively. McDonald-Kreitman tests (Betrán and Long 2003 and result above) reveal that recurrent positive selection explains these results instead of relaxation of selection. In *D. ananassae* and *D. atripex,* the inability to reject the null hypothesis of neutral evolution using Tajima's D statistic revealed that despite the recurrent positive selection there is no evidence of a recent selective sweep.

From these data, it can be inferred that *Dntf-2* retrogenes have been recurrently recruited and undergone positive selection at some point during the course of their existence in every instance that we were able to test (i.e. we have no polymorphism data for D. grimshawi and related species). With the aide of new software, it is possible to determine not only the mode of evolution acting on newly created genes but also narrow down which amino acids have been positively selected. The positively selected sites detected in the *D. melanogaster* complex reveal that although non-conservative amino acid replacements have occurred, Dntf-2r is likely capable of interaction with Ran and to some extent with nucleoporins.

A lengthier discussion about the possible selective forces that might drive the fast evolution of these genes will be presented in Chapter 4.

CHAPTER 3

EVOLUTION OF *Ran* DERIVED RETROGENES

3.1 Introduction

As introduced in Chapter 2, the selective regime a gene is undergoing or has experienced in the past can be studied using divergence data (K_A/K_S analyses), polymorphism data (Tajima's D) or a combination of both polymorphism and divergence (McDonald-Kreitman test) (Goldman and Yang, 1994; Yang, 1998; Yang et al., 2000; Yang and Neilsen, 2000; Tajima, 1989; McDonald and Kreitmant, 1991).

In this chapter, the mode of evolution of *Ran* retrogenes and parentals is studied using divergence data. *Ran-like* that is present in all the species of the *D. melanogaster* subgroup and *D. ananassae Ran* retrogene are studied in more detail. In particular, we gathered polymorphism and divergence data for both of these genes and performed the above mentioned analyses. Interestingly, recurrent positive selection has again been acting in these *Ran* duplicates.

3.2 Methods

3.2.1. PAML and HyPhy software analyses

The nucleotide sequences of the *Ran* gene in the 12 sequenced *Drosophila* species were retrieved from FlyBase (Clark et al., 2007) along with the orthologous *Ran-like* sequences in *D. melanogaster*, *D. simulans*, *D. sechellia and D. erecta*. *D. yakuba Ran-like* sequences were not included in any of the analyses below because *Ran-like* was discovered to be degenerating. This gene has a missannotated intron in FlyBase and polymorphism data reveal mostly disabled alleles. These 16 sequences were aligned with each other along with the retroposed *Ran* sequences from *D. ananassae* and *D. grimshawi* with Clustal W software

(Thompson et al., 1994). Sequences were analyzed using the CODEML software package implemented in PAML 3.15 (Yang 1997). The tree provided was (((((((((*Ran-like_D. simulans*, *Ran-like_D. sechellia*), *Ran-like_D. melanogaster*), *Ran-like_D.erecta*) ((*Ran_D. simulans*, *Ran_D. sechellia*), *Ran_D. melanogaster*)), (*Ran_D.yakuba*, *Ran_D.erecta*)), (Da_*Ran-like_D. ananassae*, *Ran_D. ananassae*)), (*Ran_D.psedoobscura*, *Ran_D.persimilis*)), *Ran_D.willistoni*), ((*Ran_D.virilis*, *Ran_D.mojavensis*), (Dg_*Ran-like_D.grimshawi*, *Ran_D.grimshawi*))) and the topology is seen in figure 3.1. Branch models were employed to determine selective pressures on the parental and retroposed sequences by calculation of ω ratios for each branch (Yang 1998). The free-ratio model allows every branch to evolve at a different rate, and the one-rate model sets all branches to evolve at the same rate. Other models allow for differing ω rates for all retroposed sequences and *Ran* branches (two-ratio model), and independent ω ratios for each retroposed sequence along with one-ratio for the *Ran* branches (four ratio model). A five-ratio model was fit to test whether the branch following duplication of *Ran* in the *D. melanogaster* complex (Node 4…Node 5, (see figure 3.1)) experienced accelerated evolution relative to subsequent branches. An eight-ratio model was created to test if *Ran* branches that have given rise to a retrogene (*D. melanogaster* complex, *D. ananassae*, and *D. grimshawi*) have evolved with differing ω ratios relative to *Ran* lineages that have not given rise to a retrogene. To resolve whether *Ran-like* branches evolve at different rates a ten-ratio model was created to assign independent ω ratios to each *Ran-like* branch in the phylogeny. Finally, multiple null models were fit to the data wherein ω is fixed at one (ω=1) in all lineages, one ω group at a time, and compared with the best statistical alternative model to detect positive/purifying selection. These were all a priori hypothesis to be tested. 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.

Next, site models of CODEML software implemented in PAML were used to uncover the possibility of positive selection acting on a few sites. Site-specific likelihood model pairs M1 (nearly neutral) and M2 (positive selection) along with M7 and M8 were applied to the sequences with the appropriate tree topology (Nielsen and Yang 1998; Yang et al. 2000). Models M1 and M7 do not allow for sites under positive selection and are compared to models M2 and M8 respectively, both of which allow for sites under positive selection. All models allow variable selective pressures among sites but fix rates among branches in the phylogeny. Model M2 is an extension of M1 by incorporation of an additional rate class $\omega_2 > 1$ (estimated from data with proportion p_2) to the two rate classes ($\omega_0 << 1$ and $\omega_1 = 1$) and proportions (p_0 and p_1) present in M1. M7 assumes a beta distribution for ω between 0 and 1 over all sites while M8 adds an additional site class ($\omega \geq 1$) with ω estimated from the data. A likelihood ratio test (LRT) was performed for both pairs by calculating two times the log likelihood values and comparing this value to a χ^2 distribution with 2 degrees of freedom. Posterior probabilities of codons under positive selection are computed in models M2 and M8 using Bayes Empirical Bayes when the LRT was significant. The tree provided for the site analyses for *Ran-like* was (((*Ran_like_D. simulans*, *Ran_like_D. sechellia*), *Ran_like_D. melanogaster*), *Ran_like_D. erecta*).

The same eighteen *Ran* and *Ran* retrogene aligned sequences along with the tree topology used for PAML branch-specific analyses were uploaded to the HyPhy software package available at http://www.datamonkey.org. The GA branch model was run to uncover when selection occurred in the phylogeny. The models tested in the GA branch test are similar to the branch-site models and the branch-specific models employed in the PAML software package. As in the branch-site models, branch lengths and substitution rates are calculated by maximum likelihood for the phylogeny and held constant while site-to-site rate variation is calculated for codons in the sequence alignment. Finally, individual branches are automatically partitioned into different discrete classes with each class having independent ω ratios until a general optimized model is found. While the PAML branch-specific models bin branches

specified by the user *a priori*, the GA branch models bin branches into an increasing number of ω ratio classes until the most probable model is located (Kosakovsky Pond and Frost, 2005).

Ran-like sequences and tree topology (((*Ran-like_D. simulans*, *Ran-like_D. sechellia*), *Ran-like_D. melanogaster*), *Ran-like_D.erecta*) were uploaded to the HyPhy software package available at http://www.datamonkey.org. REL analysis was performed in an attempt to detect positively selected codons in the *Dntf-2r* phylogeny at a Bayes factor (P value is roughly equivalent to 1/Bayes factor) threshold of 50 which corresponds to a small P value. FEL analysis was also done (using the same sequences and tree topology as REL analysis) to detect positively selected codons using a cutoff value of P<0.1 (Kosakovsky Pond and Frost, 2005)

3.2.2. Strains and sequencing

Genomic DNA extractions were performed on single fly using Puregene kit. *Ran-like* was PCR amplified from genomic DNA from twelve *D. melanogaster* flies from different Zimbabwe strains (ZH13, ZH18, ZH19, ZH20, ZH21, ZH23, ZH26, ZH27, ZH28, ZH29, ZH32, and ZH40; (Hollocher et al. 1997), nine *D. simulans* flies from different Madagascar strains (M1, M4, M5, M24, M37, M50, M242, M252, and M258) and eight *D. yakuba* strains (Tai6, Tai15, Tai18, Tai21, Tai26, Tai27, Tai30, and Tai159,) collected by Daniel Lachaise in the Taï forest in Ivory Coast in 1981. These strains were kindly provided by the Wu, the Aquadro and the Long laboratories respectively. Oligoprimers 5'-CTGGCAGGATAGGTTCAATAC-3' and 5'-CAAAGATCATCGTTGCAC-3' were used for amplification in *D. melanogaster.* Primers 5'GCTGGCGGGATAAGTTC3' and 5'CCATGGGCACGAAGTAAG3' were used for amplification in *D. simulans*. For *D. yakuba,* oligoprimers 5'ATTACACAAGCCGCTCC3' and 5'ACGCAGAAGGGGAAAAG3' were used.

Genomic DNA extractions were performed on single fly using Puregene kit. *Ran* retroposed sequences were PCR amplified from genomic DNA from ten *D. ananassae* strains (14024-0371.16, 14024-0371.17, 14024-0371.18, 14024-0371.25, 14024-0371.30, 14024-

0371.31, 14024-0371.32, 14024-0371.33, 14024-0371.34, and 14024-0371.35), and four *D. atripex* strains (14024-0361.00, 14024-0361.01, 14024-0361.02, and 14024-0361.03). These strains were acquired from UC San Diego *Drosophila* stock center. Oligoprimers 5'- CAA TCT CCT CGT GCA GAC G -3' and 5'- CGG AGT GTC CAA TTT GTC G -3' were used for amplification of the retroposed *Da_Ran-like* gene in all *D.atripex*, while oligoprimers 5'- CAA TCT CCT CGT GCA GAC G -3 and 5'- GCA ACG CCA CTT TCG TG -3' were used to amplify *Da_Ran-like* in *D. ananassae*. PCR products were sequenced from both strands using an automated DNA sequencer using fluorescent HiDi terminators. PCR products of heterozygous individuals for this gene were cloned and a clone was sequenced to establish the haplotypes. McDonald-Kreitman test (McDonald and Kreitman 1991) was performed using the polymorphism data obtained for the retroposed *Ran* sequence in *D. ananassae*, and *D. atripex*. Sequenced products were aligned using Clustal W (Thompson et al, 1994) and imported into DnaSP 4.0 (Rozas et al. 2003) to perform the McDonald-Kreitman test and for calculation of Tajima's D statistic.

3.3 Results

3.3.1. PAML and HyPhy software analyses

To understand the mode of evolution of *Ran* and its retrogenes we performed sequence analyses as described in the methods section. Log likelihood values and maximum likelihood estimates of K_A/K_S ratios for the branches in the *Ran, Ran-like* and other *Ran* retrogenes phylogeny using PAML are given in supplementary table 3. A free-ratio model (data not shown) was fitted (l = -4410.177751, *p* = 69) and compared to the one-ratio model. The free-ratio model gave a much better fit (P = 0) to the data due to differing K_A/K_S ratios along the branches of the tree. The one-ratio model was then compared to a two-ratio model (P = 0) showing a 43 fold increase in the rate of evolution in the *Ran* retrogene lineages when compared to the *Ran* branches (i.e. 0.1793 for retrogene *vs.* 0.0042 for parental gene). Next, a four-ratio model was fitted to the data to allow differing rates of evolution for the branches which correspond to the

three recurrent recruitments of *Ran* retrogenes. This model shows accelerated evolution of *Ran-like* in the *D. melanogaster* subgroup relative to all other branches in the tree when compared to the two-ratio model (P = 0). Other retrogene lineages are evolving much faster that the parental genes as well (9 to 10 times faster). Lastly we wanted to determine the mode of evolution in the *D. melanogaster* subgroup immediately after duplication of *Ran*. The five-ratio model is an extension of the four-ratio model and allows the branch after duplication (Node 4…Node 5 in figure 3.2) to evolve under a separate K_A/K_S ratio. This five-ratio model is significantly better than the previous four-ratio model ($P = 1.8873 \times 10^{-15}$). Interestingly, we found that immediately after duplication of *Ran* in the *D. melanogaster* subgroup purifying selection ($K_A/K_S = 0.0249$) was acting on the newly retroposed gene but it was still evolving ~6 times faster than the parental. Both relaxation of selection or positive selection could explain this general increase in evolutionary rate in the retrogenes. In addition, this five-ratio model shows a marked increase of K_A/K_S (0.7023) in the *D. melanogaster* subgroup when the branches in the different lineages are allowed to evolve at a single different rate. Models that consider a different rate for the parental genes after duplication (eight ratio) and individual *Ran-like* branches in the *D. melanogaster* subgroup (ten ratio) were significantly worse at fitting the data than the five-ratio model (data not shown). All ω ratios in the five-ratio model are significantly smaller than one after five null models were fit (Table 3.1 only data for *Ran-like* is shown). Again, both relaxation of constraint or positive selection could explain this enormous increase in evolutionary rate in the *Ran-like* lineages. Additional analyses (see below) reveal that positive selection has acted in some of these lineages.

Table 3.1 Results of PAML *Ran-like* branch-specific analyses.

Model	l	p	$\hat{w}_{Ran}$	$\hat{w}_{Ran\text{-}like}$	$\hat{w}_{Ran\text{-}like\_dup}$	$\hat{w}_{Da\_Ran\text{-}like}$	$\hat{w}_{Dg\_Ran\text{-}like}$
One-ratio	-4610.416	36	0.0571	0.0571	0.0571	0.0571	0.0571
Two-ratio	-4492.862	37	0.0042	0.1793	0.1793	0.1793	0.1793
Four-ratio	-4452.454	39	0.0043	0.3593	0.3593	0.0395	0.0349
Five-ratio	-4420.883	40	0.0044	0.7023	0.0249	0.0408	0.0348
Five-ratio null	-4422.926	39	0.0044	1	0.025	0.0409	0.035

. l refers to log likelihood values. p is the number of parameters estimated in the model. $\hat{w}_{Ran}$ is K_A/K_S ratio for all ran genes. $\hat{w}_{Ran\text{-}like}$ is the K_A/K_S ratio for the melanogaster subgroup minus the branch immediately following duplication of *Ran-like*. $\hat{w}_{Ran\text{-}like\_dup}$ is the K_A/K_S ratio for the branch (Node 4….Node 5 in figure 3.1) immediately following duplication in the melanogaster subgroup. $\hat{w}_{Da\_Ran\text{-}like}$ and $\hat{w}_{Dg\_Ran\text{-}like}$ the K_A/K_S ratio for the retroposed sequence in *D. ananassae* and *D. grimshawi* respectively.

The results of the GA-branch analysis are consistent with results from PAML branch-specific models. Because branches of interest are not specified a priori, this analysis is likely to bin branches into ω bins that more accurately reflect selective pressures. For example, some *Ran-like* branches in the *D. melanogaster* subgroup have ω ratios exceeding 1 (seen in red) indicating positive selection on those branches while other branches nearby have lower ω values than predicted under the PAML branch-specific models. Overall, the same trend can be seen in both GA-branch and PAML branch-specific analyses where retrogenes have experienced accelerated evolution relative to the non-retroposed genes. It is interesting to note that the branch after the *Ran* duplication occurred in the *D. melanogaster* subgroup (Node 4…Node 5, figure 3.2) has a much lower ω ratio than the subsequent branches (consistent with PAML analysis), while the previous branch (Node 3…Node 4, figure 3.2) may have been under positive selection. The timing suggests that around the time of *Ran* duplication in the *D. melanogaster* subgroup (and duplication in *D. ananassae*, Node 3…Node 21), there was likely a period of accelerated evolution which is consistent with the original hypothesis of immediate

positive selection after gene duplication. This result was not uncovered in the PAML analysis because branches to be tested were specified a priori.

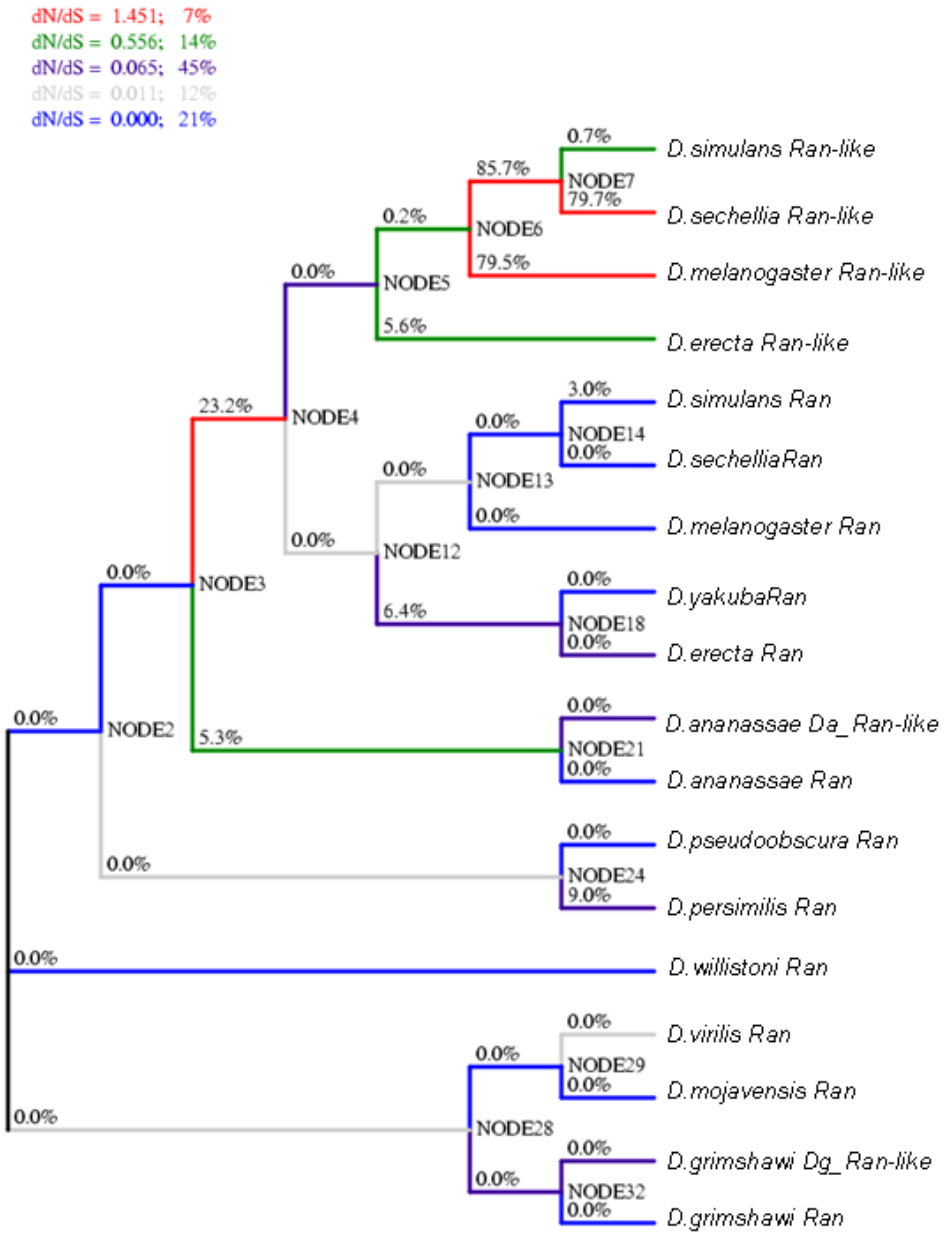

Figure 3.1 Results of GA-branch *Ran-like* analysis from HyPhy software package. Percentages over branches are the probability of that branch being under positive selection.

Site models (pairs M1 and M2, and M7 and M8) were also fitted to the data to test for positive selection acting on particular sites of *Ran-like* (See Materials and Methods). The likelihood ratio test between site models M1 and M2 was statistically significant ($2\Delta l = 8.814$, d.f. = 2; $P = 0.012$) as was the comparison between models M7 and M8 ($2\Delta l = 8.885$, d.f. = 2; $P = 0.0118$) which is indicative of positive selection acting on *Ran-like*. Furthermore, these results reveal that models M2 and M8 (allowing for positively selected sites) significantly fit the data better than models M1 and M7. Site models M2 and M8 estimated that 36.4% of sites (table 3.2) in the *Ran-like* alignment experienced positive selection ($\omega = 2.52$). Codons that are most likely under positive selection as revealed by Bayes empirical Bayes analyses (i.e. posterior probabilities ≥ 0.95%) are shown in table 3.3 and figure 3.1. Random effects likelihood (REL) analyses using Hyphy package detected 14 sites (Table 3.3 and Figure 2) with a Bayes factor bigger than 50 ($P<0.02$) that have likely been under positive selection. The more stringent analyses (i.e. FEL) detected only one codon (131) as being likely under positive selection ($P = 0.09733$).

Table 3.2 Results of PAML *Ran-like* site-model analyses.

Model	p	l	Parameter Estimates	Postively Selected Sites
M1: Neutral	7	-1704.602	$p_0 = 0.385$ $p_1 = 0.615$	NA
M2: Selection	9	-1700.195	$p_0 = 0.635$, $p_1 = 0$, **$p2 = 0.365$, $w2 = 2.51$**	Many (see table 3.3)
M7: beta	7	-1704.638	$p = 0.00753$, $q = 0.005$	NA
M8: beta + ω	9	-1700.196	$p = 25$, $q = 99$, **$p_1 = 0.364$, $\omega_1 = 2.52$**	Many (see table 3.3)

ℓ represents the log likelihood value for a particular model and p refers to the number of parameters estimated in the model.

The sites under positive selection were mapped to an alignment of *Ran* and *Ran-like* from *D. melanogaster*, *D. simulans*, and *D. sechellia*. Because RanGTP/RanGTP has been crystallized with multiple interacting proteins, it is known which amino acids are responsible for

interacting with each protein. Some of the positively selected codons have occurred in a region responsible for interaction with RanGEF (codons 94 and 95). Amino Acid 140 is has been shown to interact with Importin beta, while amino acid 82 lies between two amino acids that also interact with Importin β. These changes are not thought of as conservative changes and could alter the ability of Ran to interface with RanGEF and Importin β. All other positively selected sites have occurred in regions not known to have any interaction with other proteins.

Table 3.3 Positively selected codons indicated by PAML site specific analyses and REL analysis. Codon amino acids are relative to *D. simulans*.

Codon	M8 BEB Probability	M2 BEB Probability	REL Bayes Factor
10 T	0.906	0.842	44.4
48 N	0.911	0.827	**572.1**
49 H	**0.955**	0.904	**558.1**
58 V	0.810	0.709	**52.2**
81 I	0.748	0.637	**53.1**
83 S	0.908	0.848	49.9
92 T	0.785	0.684	**54.2**
93 A	0.913	0.831	**598.3**
94 K	0.825	0.729	**53.1**
95 A	**0.958**	0.911	**469.5**
131 S	0.939	0.875	**583.9**
140 R	**0.960**	0.913	**467.6**
200 F	**0.985**	**0.964**	**3574.9**
202 D	0.888	0.793	**443.9**
207 Y	0.842	0.753	**55.3**
215 F	**0.921**	0.844	**389.1**

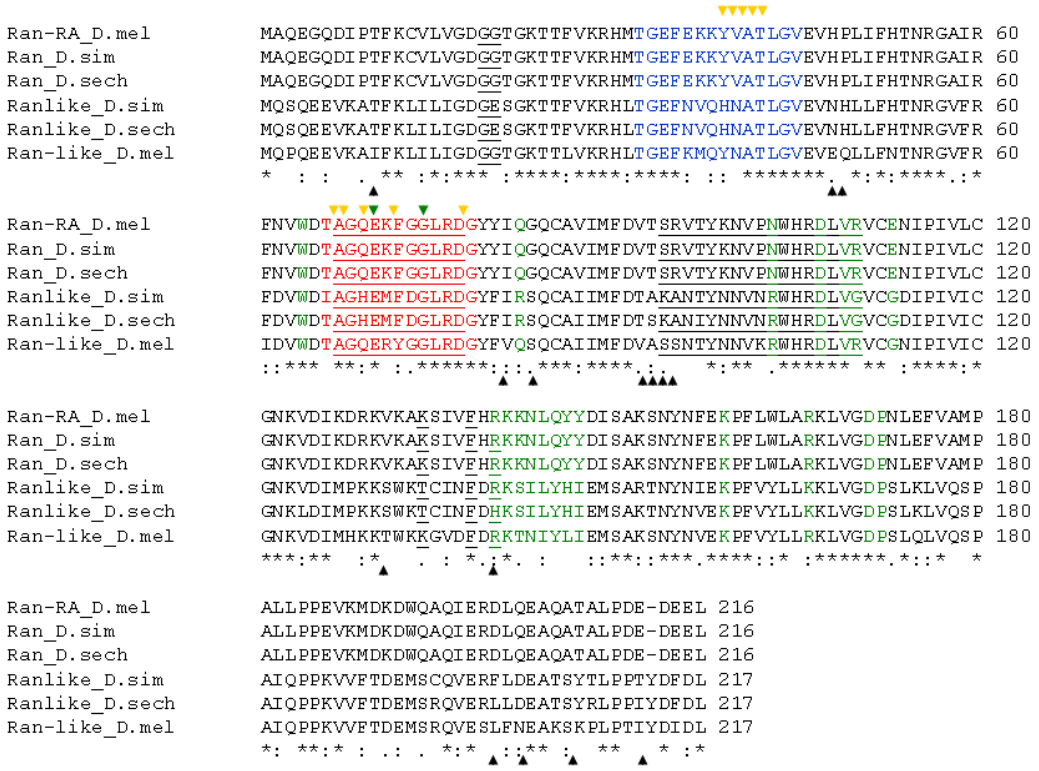

Figure 3.2 *Ran* and *Ran-like* amino acid alignment. Black triangles indicate positively selected amino acids. Gold triangles indicate residues that interact with Dntf-2. Blue residues indicate switch I. Red residues indicate switch II. Underlines residues interact with RCC1. Green residues and green triangles indicate amino acids that interact with importin beta.

3.3.2. Polymorphism data

Polymorphism data in *D. melanogaster, D. simulans* and *D. yakuba* for *Ran-like* was obtained as discussed in section 3.2.2. Fifteen alleles were sequenced for *D. melanogaster*, nine alleles were sequenced for *D. simulans*, and eight alleles were sequenced for *D. yakuba*. The McDonald-Kreitman test was performed for *Ran-like* using polymorphism data for *D. simulans* and *D. melanogaster*. Results appear in table 3.4. A significantly higher ratio of replacements per silent substitutions was found when compared to the ratio of replacements per silent polymorphisms (G (with Williams correction) = 4.062, P = 0.0438). The modified McDonald-Kreitman test excluding unpreferred changes (those unlikely to be fixed in the

population due to codon bias) was also significant (G (with Williams correction) = 4.028, *P* = 0.0447). These data supports recurrent positive selection acting on *Ran-like* in *D. melanogaster* and *D. simulans*. Lineage specific McDonald-Kreitman test performed for *Ran-like* in *D. melanogaster* and *D. simulans* (polarized with *Ran-like* from *D.erecta*) were not significant (data not shown). The only statistical significance occurs when all the data is together.

Table 3.4 McDonald-Kreitman test for *Ran-like* for *D. melanogaster* and *D. simulans*.

	Fixed	**Polymorphic**
Replacement	47	21
Synonymous	12 (2)	14 (5)

G (with Williams correction) = 4.062, *P* = 0.0438*. Values in parentheses correspond to the changes to preferred codons used for the modified McDonald-Kreitman test: G (with Williams correction) = 4.028, *P* = 0.0447*.

As introduced in the methods, *D. yakuba* polymorphism data were left out of the McDonald-Kreitman analysis because most *Ran-like* alleles were found to have deletions in the coding region (alignment can be seen in Appendix B). All the deletions result in changes in frame and/or premature stop codons indicating nonfunctionalization of the retrogene in this species. This gene loss is interesting for the discussion of the function of these retrogenes and will be revisited in Chapter 4.

Polymorphism data was also obtained in *D. ananassae* and *D. atripex* for *Da_Ran-like* (Table 3.5). Thirteen alleles were obtained for *D. ananassae* and five alleles were sequenced for *D. atripex.* The McDonald-Kreitman test was performed and revealed a significantly higher ratio of replacements per substitutions was found when compared to the ratio of replacements per silent polymorphisms (Fisher's exact test *P* (two tailed) = 0.003124**; Table 3.5). This result is compatible with recurrent positive selection occurring in these lineages.

Of sequences used for the McDonald-Kreitman tests, one allele per isoline was used to calculate Tajima's D statistic for *D. melanogaster, D. simulans, D. ananassae* and *D.atripex*. For *D. melanogaster* and *D. simulans* Tajima's D was calculated to be -1.09956 and -1.41903 respectively. Negative values obtained for the D statistic stem from a high levels of polymorphic sites relative to the average number of polymorphisms in isoline comparisons (Tajima, 1989). However, in both instances there remains an inability to reject the null hypothesis of neutrally evolving genes. The D statistics for the *D. ananassae* and *D.atripex* polymorphism data sets are 0.23932 and 2.01187 respectively. Both D statistics are not statistically significant and these positive values are generally indicative of population structure or a population under balancing selection (Tajima, 1989). The D statistics may not be completely appropriate as the sampled flies come from very different populations and population structure can highly influence the statistic.

Table 3.5 McDonald-Kreitman test for the retroposed *Ran* genes (*Da_Ran-like*) in *D. ananassae* and *D.atripex*.

	Fixed	**Polymorphic**
Replacement	15	0
Synonymous	33	21

Fisher's exact test *P* (two tailed) = 0.003124**. The G statistic cannot be calculated when a zero occurs in the contingency table.

3.4 Discussion

An extensive analysis of *Ran* and retroposed sequences evolution using 12 sequenced Drosophila genomes has been performed. PAML branch-specific Ka/Ks analyses have revealed a 160x and 9x increase in the evolutionary rate of retroposed *Ran* genes in the *D. melanogaster* subgroup and the *D. ananassae* lineage respectively, while site-specific models and GA-branch

analyses indicate positive selection occurring at some points during the evolutionary history of *Ran-like* in the *D. melanogaster* subgroup. Interestingly, despite this acceleration ratios were still smaller than one revealing purifying selection. Furthermore, McDonald-Kreitman tests (result above) reveal that recurrent positive selection has apparently occurred on the recurrently retroposed *Ran* genes. In all species with polymorphism data available, the inability to reject the null hypothesis of neutral evolution using Tajima's D statistic revealed that despite the recurrent positive selection there is no evidence of a recent selective sweep.

From these data, it can be inferred that *Ran* retrogenes have been recurrently recruited and undergone positive selection at several points during the course of their existence. With the aide of newly developed software, it was possible to determine not only the mode of evolution acting on newly created genes but also narrow down which amino acids have been positively selected. Some information can be gleaned from the 16 codons found to be under positive selection. Many of these sites are in a region of the protein that is not known to interact with other proteins, however there is some evidence (nonconservative amino acid replacements) that the newly evolved *Ran* retrogenes in the *D. melanogaster* subgroup may be losing some ability to interface with RanGEF and Importin β underpositive selection. The consequences of these changes will be discussed in greater detail in chapter 4.

CHAPTER 4

DO THESE RETROGENES HAVE A ROLE IN MEIOTIC DRIVE?

4.1 Inferences regarding protein interactions

Heterospecific complexes formed between rat Ntf-2 and canine Ran have been crystallized and their mutual interactions are known (Berman et al. 2002). Additional crystal structures reveal in detail the amino acids and the domains involved in the interacting interfaces of Ntf-2 and Ran with other proteins (Bullock et al. 1996; Stewart et al. 1998; Vetter et al. 1999; Renault et al. 2001; Seewald et al. 2002; Matsuura and Stewart 2004; Isgro and Schulten; 2007). Ntf-2 is a dimer that interacts strongly with two and possibly three molecules of Ran bound to GDP. Ntf-2 also interacts with nucleoporins during transport to the nucleus. See figure 2.2 for details about the particular residues involved. Ran (switch I and II) interacts with Ntf-2 in its GDP bound form during transport to the nucleus, with RanGEF in its GDP bound form in the nucleus, with Importin β in its GTP bound form during its transport out of the nucleus, with exportins in its GTP bound form in complexes that are transported out of the nucleus, and with RanGAP in its GTP bound form in the cytoplasm (see figure 3.2 for details). RanGAP is localized in the cytoplasm and hydrolyzes RanGTP into Ran GDP. RanGEF (also known as RCC1) localizes in the nucleus and is a factor that produces the exchange of GDP by GTP. Together all these proteins maintain a gradient of RanGDP/RanGTP that is important for protein import and export. Concentration of RanGTP is high in the nucleus and the concentration of RanGDP is high in the cytoplasm.

It has been shown that *Dntf-2* is a gene under purifying selection in Drosophila (K_A/K_S =0.0247; see table 2.1) but *Ran* is under much stronger purifying selection (K_A/K_S =0.0044; see table 3.1) likely due to the fact that it carries multiple functions as discussed above. This is also evident from the alignments of these Drosophila proteins with other animal orthologs (Figure

4.1). Alignments of *Drosophila* Dntf-2 and Ran and their retrogene proteins with the rat Ntf-2 and canine Ran respectively have been created to explore any possible changes in function in the newly duplicated proteins (alignments not shown).

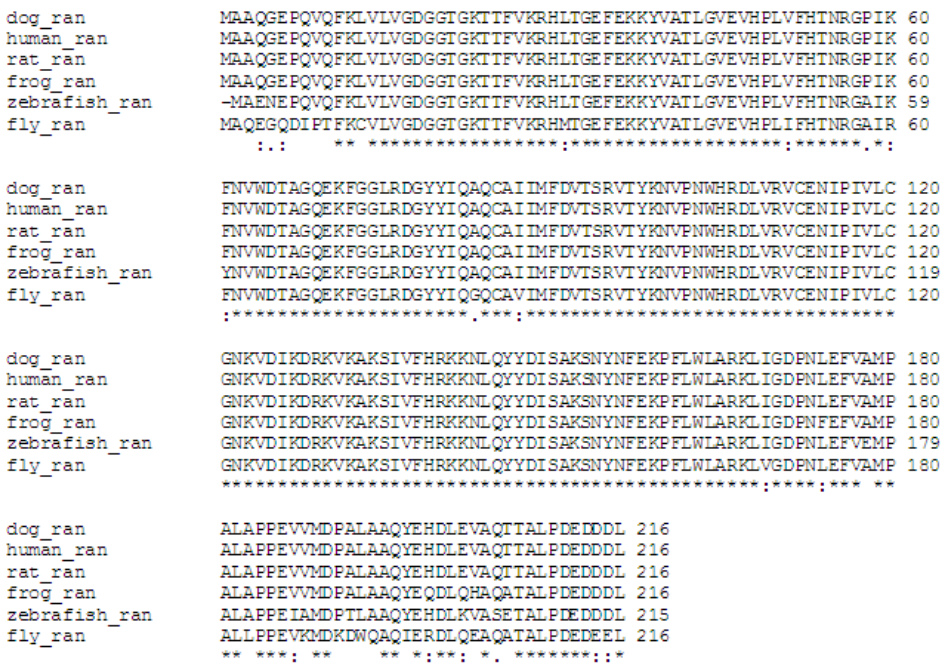

Figure 4.1 Alignment of *Ran* amino acid sequence for multiple animal species.

Based on amino acid sequence conservation we can see that the amino acids of Dntf-2 involved in interaction with RanGDP and the FxFG repeats of the nucleoporins in the nuclear pore complex are completely conserved or have conservative changes between *Dntf-2* and *Dntf-2* retrogenes in the three lineages (figure 2.2). Even in *D. ananassae Dntf-2* retrogene that is the most divergent from *Dntf-2*, most residues important for function seem to be conserved suggesting that the *Dntf-2*-retrogenes still encode proteins that can function as transport proteins and can carry molecules of RanGDP across the nuclear membrane. *Dg_Dntf-2r* and

Dntf-2 in *D. grimshawi* are not highly divergent as seen in figure 4.3 suggesting that the retroposed copy may still be able to carry out the functions of the parental gene.

```
Da_Dntf2-r_D.ana  MPLNPHYEPMGQEFVKQYYVIFDNPATRALTATFFSHNDSFMTFEGEQVLGYYKIFEKVK 60
Dntf2_D.ana       MSLNPQYEDIGKGFVQQYYAIFDDPANRANVVNFYSATDSFMTFEGHQIQGAPKILEKVQ 60
                  *.***:** :*: **:***.***:**.** ...*:* .********.*: *  **:***:
Da_Dntf2-r_D.ana  SLPFQKVNRTLTNVDCQPTGDGGILMSVLGRLQCDDDPSLSFSQIFVLKPDTSPNAYYLS 120
Dntf2_D.ana       SLSFQKITRVITTVDSQPTFDGGVLINVLGRLQCDEDPPHAFSQVFVLKANAG--TFFVA 118
                  **.***:.*.:*.**.*** ***:*:.********:**. :***:****.:::.  :::::
Da_Dntf2-r_D.ana  HDIFRLNIHDTE 132
Dntf2_D.ana       HDIFRLNIHNSA 130
                  *********::
```

Figure 4.2 Amino Acid alignment of *Dntf-2* and *Da_Dntf-2r* in *D. ananassae*. Red residues interact with RanGDP.

```
Dntf2_retgrim     MAINPQYEAVGKGFVQQYYAIFDDPANRANVVNFYSTTDSFMTFEGHQIQGAPKILEKVQ 60
Dntf2_D.grim      MAINPQYEDIGKGFVQQYYAIFDDPANRANVVNFYSATDSFMTFEGHQIQGAPKILEKVQ 60
                  **:***** :************************:********************
Dntf2_retgrim     SLSFQKINRIITTIDSQPTFDGGVLINVLGRLQCDEDPPHSYSQVFLLKANAGSFFVAHD 120
Dntf2_D.grim      SLSFQKISRVITTVDSQPTFDGGVLINVLGRVQCDDDPPLAFSQVFVLKANAGTYYVAHD 120
                  *******.*:***:*****************:***:*** ::****:******:::****
Dntf2_retgrim     IFRLNIHNSA 130
Dntf2_D.grim      IFRLNIHNSA 130
                  **********
```

Figure 4.3 Amino acid alignment of *Dntf-2* and *Dg_Dntf-2r* (retgrim). Red residues interact with RanGDP.

Similar comparisons between *Ran* retrogenes and parental amino acid sequences (alignments not shown) show that the proteins encoded by the retrogenes in *D. ananassae* and *D. grimshawi* lineages show very little divergence from their parental genes. This trend can also be seen by very low K_A/K_S values. Most of the amino acids that are important for function are conserved or show conservative changes suggesting similar functions between Ran and its retrogenes in these lineages. However, Ran-like in the *D. melanogaster* subgroup is highly diverged (Figure 3.2). As commented above, the interaction with Dntf-2 is likely still present

because most of the interacting amino acids are conserved or show conservative changes. It is also likely that Ran-like still interacts with RanGap although many of the divergent sites (Figure 3.2) cluster in a region of Ran-like/RanGap interaction or close and other interacting sites have changed although they were not detected as being under positive selection (for example codon 128 has a different amino acid in Ran-like). However these changes are classify at this point as conservative because they do not seem to produce major conflicts in the interactions. RanGap has been shown to be under positive selection itself (Presgraves 2007) and produce duplicate genes (i.e. Sd; (Kusano et al. 2003)) and interactions with these changing protein or duplicates might explain the changes. However, all other interactions seem to be weaker or absent. The analyses below focus on *D. melanogaster* Ran-like but similar conclusions apply to the other Ran-like lineages analyzed. Regarding the interaction with Importin β, Ran-like proteins seem to have reduced overall charge interactions through partial or complete changes in charge (E113G [disrupts hydrogen bond], K142T [in the basic patch], Y146L, and Y147I). The C terminal end (DEDEEL) that is known to stabilize RanGDP (Seewald et al. 2002) has diverged greatly. It is known that in the absence of this end, the RanGAP mediated hydrolysis of RanGTP to RanGDP is accelerated (Seewald et al. 2002) and the exchange of GDP to GTP catalyzed by RanGEF is also accelerated (Richards et al. 1995). The C terminal end is also required for the efficient binding of Ran to several of the Ran binding proteins. Such binding is required for proper function of Ran (Richards et al. 1995) but it is highly diverged in Ran-like. RanGEF and exportins may also have a weaker interaction with Ran-like in *D. melanogaster* because residue 37 involved in exportin interaction has change dramatically in charge and size (K37M). Ran-like residues involved in RanGEF interaction have lost charge (partially or completely) or hydrophobicity (R95S, and V96N). Residue 95 is likely under positive selection (Table 3.3 and Figure 3.2).

The above analysis seems to indicate that the *D. melanogaster* subgroup Ran-like protein has retained Dntf-2 and RanGap interfaces while losing, or at least diminishing, all other

protein-protein interfaces. The presence of Dntf-2 and RanGap interfaces on Ran-like suggests that the Ran-like protein can exist in the RanGDP form and can be carried into the nucleus by Dntf-2r where the Ran-like RanGDP could be converted to RanGTP. The importin β interface on Ran-like, however, is diminished. As a result, Ran-GTP will not be transported out of the nucleus by importin β and importin β will have a diminished capacity to release cargo upon nuclear entry. Similarly, export of RanGTP by the exportin complex may be reduced. Additionally, the loss of Ran-like's C terminal residues suggest that hydrolysis of RanGTP to RanGDP might be accelerated in the presence of RanGAP. Other changes in Ran-like may affect the exchange of GDP to GTP by RanGEF. The binding of several lesser-known Ran binding proteins may also be affected. From this data it seems that Ran-like can not completely replace Ran in the male germline in the *D. melanogaster* subgroup species where it is still present.

4.2 Are X to autosome nuclear transport retrogenes recurrently recruited for segregation distortion in *Drosophila*?

When convergent evolution occurs, it reveals how the same and strong selective pressures are acting in different lineages. A good example of convergence that involves recurrent emergence of retrogenes of the same parental gene that are believed to have acquired the same male function is Utp14 in mammals (Bradley et al. 2004). In this case, the authors concluded that X inactivation was the selective force driving this convergent recruitment of retrogenes.

Given that there is an estimated rate of retrogene formation of 0.51 retrogenes per My (Bai et al. 2007). If we consider 13,600 genes in the Drosophila genome (Adams et al. 2000), we have a rate of 3.75×10^{-5} per gene per million years. This is the probability of seeing a duplicate for any particular gene. In the whole tree analyzed by Bai et al. (2007) that has ~400 My in total (Tamura et al. 2004), the probability that a particular gene (i.e. Dntf-2 or Ran) generates a duplicate is 0.015 ($3.75 \times 10^{-5} \times 400$). Then the probability of three Ran (or Dntf-2)

duplications in this tree is 0.015^3 ($3.375x10^{-6}$) and the probability that the particular gene that interacts with the parental gene of this retrogene also duplicates is 0.015^3 $x0.015^3$ ($1.14x10^{-11}$). When we multiply this probability by 13,600 genes in the genome to which this could happen, the probability of this event (i.e. that any gene and its closer interacting partner produce retroposed copies three times in the tree of the 12 Drosophila species) is very small ($P=1.14x10^{-11}$ x 13,600 $=1.55x10^{-7}$). This does not account for the duplication being observed in the same lineages. Accounting for this would make the probability even smaller. This small probability together with the rest of data showing that these genes are X to autosome duplications, evolve fast (in several instances under positive selection) and are mainly male germline biased support a strong selective pressure in the origin of these genes.

Three selective hypotheses have been outlined in the introduction to explain the recurrent duplication of *Dntf-2* and *Ran*: male germline X inactivation, sexual antagonism and meiotic drive. It has previously been suggested that meiotic drive can be a powerful force in shaping genes and genomes because it generates a genomic conflict that leads to an arms race (Burt and Trivers 2006). Genes involved in the drive or its suppression will evolve under recurrent positive selection (Presgraves 2007). However, male germline X inactivation, and sexual antagonism can also be strong selective forces (Betran et al. 2002; Ranz et al. 2003; Wu an Xu 2003; Betran et al. 2004; Bradley et al. 2004; Emerson et al. 2004; Bai et al. 2007) and male germline genes are known to evolve under positive selection (Haerty et al. 2007). In addition to these three hypotheses, new genes can also originate in male germline and carry new unknown functions.

Reasoning dictates that the data (i.e. the recurrent origination, male biased expression of the retrogenes, positive selection acting on them, pseudogenization of the genes in some lineages (*D. yakuba*) and loss of functions of some protein domains in some duplicates) do not consistently fit the X inactivation or sexual antagonism hypotheses. In both hypotheses the gene is recruited for a previously existing important male function. It is expected that the new

gene become specialized for the new function (by fixation of favorable random mutations through natural selection) before coming under purifying selection. If the new copy performs better than the parental gene, it is not expected to be lost or lose functions. It can be concluded that a novel dispensable function hypothesis (possibly segregation distortion) explains the data better, and that the loss of *Ran-like* can be explained by the putative disappearance of the selective pressure (i.e. segregation distortion system/s) in that lineage. In addition, the loss of particular functions that have occurred in some of these new proteins (i.e. *Ran-like* in *D. melanogaster* subgroup) is also well explained by a segregation distortion role and a selective pressure to differentiate from the parental gene. In addition, a Ph.D. student in the lab (Mansi Motiwale) has acquired a *Dntf-2r* knockout (a *D. melanogaster* P-element insertion line (EY05573)) that produces no *Dntf-2r* transcript and does not show obvious male fertility effects further supporting the dispensability of the gene (unpublished data). However, it remains possible that adaptive forces are different at different times and/or in different lineages and all possible hypotheses might have some relevance as they are not mutually exclusive.

Therefore, it is hypothesized that *Dntf-2* or *Ran* duplicates might fix in the populations if there are meiotic drive system disturbing the Ran gradient because they would probably act initially in male germline as additional doses of the parental genes. It appears that either one of them would increase the amount of RanGDP available in the nucleus to be transformed to RanGTP. This action would be compensatory to any gradient distortion as observed in previous experiments for Ran (Kusano et al. 2002). Subsequently, alleles of both genes might drive (*Dntf-2r*) or compensate (*Ran-like*) and increase or decrease in frequency depending on their linkage with respect to driving systems. This would explain the positive selection we observe in *Ran-like, Da_Ran-like,* and *Da_Ntf-2r* and has been reported for *Dntf-2r* (Betrán and Long 2003). From PAML and HyPhy analysis it has been determined which residues of *Ran-like* are likely under positive selection. Knowing the functional importance of wild type Ran, (Ciciarello et al. 2007; Clarke and Zhang 2008) it can be assumed that *Ran-like* encoded protein may have

been under positive selection to lose some of the functions so as not to interfere with the proper functioning of the cell.

Seemingly, the recurrent recruitment of *Dntf-2* and *Ran* retrogenes is likely another example of convergent evolution under the same selective pressures since they mostly are male bias in expression (see Chapter 1 and data not shown produced by Javier Río) and may have all been involved in segregation distortion functions. Experiments are being carried out to reveal the potential role of *Dntf-2r* and *Ran-like* in segregation distortion in *D. melanogaster* using the SD system. This will be done by acquiring knockout lines with P element insertions, large chromosomal deletions wherein a gene of interest has been lost, or by RNAi knockdown.

Another possible arms race these genes could be involved with is the battle against viruses and/or retrotransposable elements that need to enter the nucleus (Tang and Presgraves 2009). In their struggle to be passed to the next generation these elements must gain access to the cell nucleus where they can be reverse transcribed into the host genome. Nuclear transport gene products could also be involved in selectively allowing certain harmless molecules into the nucleus while restricting potentially hazardous RNAs. However we currently envision that this struggle would be similar in male and female germline and not only male germline.

However, as stated above, it remains possible that adaptive forces are different at different times and/or in different lineages and that the other hypotheses (X inactivation, sexual antagonism or new functions not related to meiotic drive or retroelements) remain of relevance at particular times or in some of the lineages.

APPENDIX A

POLYMORPHISM DATA

Da_Ran-like polymorphism haplotypes.

	nt.	8	19	20	57	69	85	93	96	105	108	111	114	120	126	135	153	162	174	189	234	244	246	247	249	264	297	303	318	336	341	343	354	363	390	421	423	430	432	438	441	447	453	459	465	490	498	502	513	516	517	531	546	552	558	564	571	573	578	581	591	597	605	612	618	625	627	636	648
D.atr 00_A		C	T	C	A	A	A	T	T	C	G	G	G	C	C	A	A	T	C	A	C	C	A	A	A	T	T	G	T	G	T	G	A	G	G	A	A	G	C	G	T	T	T	C	T	T	C	C	A	C	C	C	T	C	T	T	G	A	G	G	G	A	T	G	C	T	T	T	A	T	C
D.atr 00_B, 02		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.
D.atr 01 , 03		.	.	.	.	.	.	.	.	.	A	.	A	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	T
D.ana 16, 17, 35		G	G	G	T	G	C	C	.	T	.	.	.	G	T	C	T	C	.	G	A	G	G	G	G	A	.	G	C	C	C	G	A	A	G	C	.	T	.	C	C	.	.	C	C	T	A	T	T	T	C	T	C	C	A	C	C	C	T	T	G	.	T	.	A	C	G	.	.
D.ana 18_A		G	G	G	T	G	C	C	.	T	.	.	.	G	T	C	T	C	.	G	A	G	G	G	G	A	.	G	C	C	C	G	A	A	G	C	.	T	.	C	C	.	.	C	C	T	A	T	T	T	C	T	C	C	A	C	C	C	T	T	G	.	T	G	A	C	G	.	.
D.ana 18_B		G	G	G	T	G	C	C	.	T	.	.	.	G	T	C	T	C	.	G	A	G	G	G	G	A	A	G	C	C	C	G	A	A	G	C	.	T	.	C	C	.	.	C	C	T	A	T	T	T	C	T	C	C	A	C	C	C	T	T	G	.	T	.	A	C	G	.	.
D.ana 25		G	G	G	.	.	C	C	.	T	.	A	.	G	T	C	T	C	.	G	A	G	G	G	G	A	.	G	C	C	C	G	A	A	G	C	.	T	.	C	C	.	.	C	C	T	A	T	T	T	C	T	C	C	A	C	C	C	T	T	G	.	T	.	A	C	G	.	.
D.ana 30		G	G	G	T	G	C	C	.	T	.	.	.	G	T	C	T	C	.	G	A	G	G	G	G	A	.	G	C	C	C	G	A	A	G	C	.	T	.	C	C	.	.	C	C	T	A	T	T	T	C	T	C	C	A	C	C	C	T	T	G	.	T	.	A	C	G	C	.
D.ana 31		G	G	G	T	G	C	C	.	T	.	.	.	G	T	C	T	C	.	G	A	G	G	G	G	A	.	G	C	C	C	G	A	A	G	C	.	.	.	C	C	A	.	.	C	T	A	T	T	T	C	T	C	C	A	C	C	A	T	T	G	T	T	.	A	C	G	.	.
D.ana 32_A		G	G	G	.	.	C	C	.	T	.	A	.	G	T	C	T	C	.	G	A	G	G	G	G	A	.	G	C	C	C	G	A	A	G	C	A	.	.	C	C	A	.	.	C	T	A	T	T	T	C	T	C	C	A	C	C	A	T	T	G	T	T	.	A	C	G	.	.
D.ana 32_B		G	G	G	.	.	C	C	C	T	.	.	.	G	T	C	T	C	T	G	A	G	G	G	G	A	A	G	C	C	C	G	A	A	G	C	.	.	A	C	C	A	.	.	C	T	A	T	T	T	C	T	C	C	A	C	C	A	T	T	G	T	T	.	A	C	G	.	.
D.ana 33_A		G	G	G	T	.	C	C	.	T	.	.	.	G	T	C	T	C	.	G	A	G	G	G	G	A	A	G	C	C	C	G	A	A	G	C	.	.	.	C	C	A	T	.	C	T	A	T	T	T	C	T	C	C	A	C	C	A	T	T	G	T	T	.	A	C	G	.	.
D.ana 33_B		G	G	G	.	.	C	C	.	T	.	.	.	G	T	C	T	C	.	G	A	G	G	G	G	A	A	G	C	C	C	G	A	A	G	C	.	.	.	C	C	A	.	.	C	T	A	T	T	T	C	T	C	C	A	C	C	A	T	T	G	T	T	.	A	C	G	.	.
D.ana 34		G	G	G	.	.	C	C	C	T	.	.	.	G	T	C	T	C	T	G	A	G	G	G	G	A	A	G	C	C	C	G	A	A	G	C	.	.	.	C	C	A	.	.	C	T	A	T	T	T	C	T	C	C	A	C	C	A	T	T	G	T	T	.	A	C	G	.	.

D.atr and *D.ana* refer to *D. atripex* and *D. ananassae* respectively. See text for isoline details.

Da_Dntf-2r polymorphism haplotypes.

nt.	3	7	10	16	17	18	19	20	39	40	48	52	57	63	70	71	72	75	81	84	86	90	93	96	99	102	105	114	123	127	131	135	141	147	169	170	171	174	189	190	191	195	200	204	207	210	220	221	222	232	237	240	242	254	256	264	267	276	281	282	285	287	291	292	300	303	306	314	315	316	325	326
D.atr 00	C	G	T	A	C	A	A	A	C	T	C	C	G	A	C	A	C	C	T	T	C	A	C	G	T	T	C	C	C	C	C	C	A	T	T	C	A	G	C	T	G	G	T	C	C	C	C	C	T	T	T	C	G	G	C	A	T	G	T	C	G	A	T	G	A	C	G	C	C	T	G	C	A	T	C
D.atr 01_A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	A	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D.atr 01_B, 03_A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	A	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D.atr 02_A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D.atr 02_B	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	C	.	.	.	A	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D.atr 03_B	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D.ana 16, 17, 18 25_A, 30, 35	.	C	A	C	A	G	G	A	.	G	T	A	T	T	C	T	T	A	A	.	T	.	C	.	.	A	T	T	A	G	T	C	C	T	C	C	A	C	A	A	C	A	A	G	T	C	G	G	T	A	.	G	G	G	C	C	T	T	T	.	C	T	A	C	T	A	G	A	.	.	A	A
D.ana 25_B, 33_A	.	C	A	C	A	G	G	A	.	G	T	A	T	T	C	T	T	A	C	.	T	.	C	.	C	.	.	T	A	G	T	C	C	T	C	C	A	C	A	A	C	A	A	G	T	C	G	G	T	A	.	G	G	G	C	C	T	T	T	.	C	T	A	C	T	A	G	A	.	.	A	A
D.ana 31	.	C	A	C	A	G	G	A	.	G	T	A	T	T	C	T	T	A	A	.	T	.	C	.	.	A	T	T	A	G	T	C	C	T	C	C	A	C	A	A	C	A	A	G	T	C	G	G	T	A	.	G	G	G	C	C	T	T	T	.	C	T	A	C	T	A	G	A	T	C	A	A
D.ana 32	T	C	A	C	A	G	G	A	.	G	T	A	T	T	C	T	T	A	A	.	T	.	C	.	.	.	T	T	A	G	T	C	C	T	C	C	A	C	A	A	C	T	A	G	T	C	G	G	T	A	.	G	G	G	C	C	T	T	T	.	C	T	A	C	T	A	G	A	.	.	A	A
D.ana 33_B	.	C	A	C	A	G	G	A	.	G	T	A	T	T	C	T	T	A	C	.	T	T	C	C	.	.	T	T	A	G	T	C	C	T	C	C	A	C	A	A	C	A	A	G	T	C	G	G	T	A	.	G	G	G	C	C	T	T	T	.	C	T	A	C	T	A	G	A	T	C	A	A
D.ana 34	.	C	A	C	A	G	G	A	C	G	T	A	T	T	C	T	T	A	C	.	T	.	C	.	C	.	.	T	A	G	T	C	C	T	C	C	A	C	A	A	C	A	A	G	T	C	G	G	T	A	.	G	G	G	C	C	T	T	T	.	C	T	A	C	T	A	G	A	.	.	A	A

D.atr and *D.ana* refer to *D. atripex* and *D. ananassae* respectively. See text for isoline details.

Ran-like Polymorphism haplotypes.

nt.	7	29	59	62	72	78	92	105	108	109	115	121	142	144	147	153	157	180	181	186	189	197	207	211
D. simulans																								
M1	T	C	A	G	A	T	T	C	T	G	C	G	A	C	T	T	C	C	T	A	T	C	C	A
M4	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	T	C	T	.	.
M5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	T
M24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	T	.	.	.	.
M37	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M50	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M242	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
M252	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	T	.	.	.	.
M258	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
D. melanogaster																								
ZH13,19,23A,28	C	T	G	C	C	G	.	T	G	A	T	.	G	G	A	G	A	.	A	T	.	.	G	.
ZH18,21A	C	T	G	C	C	G	.	T	G	A	T	.	G	G	A	G	A	.	A	T	.	.	G	.
ZH20A	C	T	G	C	C	G	.	T	G	A	T	.	G	G	A	G	A	.	A	T	.	.	G	.
ZH20B	C	.	G	C	C	G	A	.	G	A	T	.	G	G	A	G	A	.	A	T	.	.	G	.
ZH21B,27	C	T	G	C	C	G	.	.	G	A	T	.	G	G	A	G	A	.	A	T	.	.	G	.
ZH23B	C	T	G	C	C	G	.	.	G	A	T	.	G	G	A	G	A	.	A	T	.	.	G	.
ZH26	C	T	G	C	C	G	.	.	G	A	T	.	G	G	A	G	A	.	A	T	.	.	G	.
ZH29	C	T	G	C	C	G	.	T	G	A	T	.	G	G	A	G	A	.	A	T	.	.	G	.
ZH32	C	T	G	C	C	G	.	.	G	A	T	.	G	G	A	G	A	.	A	T	.	.	G	.
ZH40	C	T	G	C	C	G	.	T	G	A	T	.	G	G	A	G	A	.	A	T	.	.	G	.

nt.	212	215	218	234	241	245	247	252	274	275	280	281	283	285	291	305	306	309	316	317	318	328	340	349
D. simulans																								
M1	T	T	A	G	A	G	T	A	A	C	A	A	G	G	C	A	T	G	C	G	C	G	G	A
M4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	A	.	.	.	C
M5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.
M37	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.
M50	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.
M242	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M252	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M258	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D. melanogaster																								
ZH13,19,23A,28	G	A	G	.	G	A	.	.	G	T	T	C	T	A	A	.	A	A	.	.	.	A	A	.
ZH18,21A	G	A	G	.	G	A	G	.	G	T	T	C	T	A	A	.	A	A	.	.	.	A	A	.
ZH20A	G	A	G	.	G	A	G	.	G	T	T	C	T	A	A	.	A	A	.	.	.	A	A	.
ZH20B	G	A	G	.	G	A	.	.	G	T	T	C	T	A	A	.	A	A	.	.	.	A	A	.
ZH21B,27	G	A	G	.	G	A	.	.	G	T	T	C	T	A	A	.	A	A	.	.	.	A	A	.
ZH23B	G	A	G	.	G	A	G	.	G	T	T	C	T	A	A	.	A	A	.	.	.	A	A	.
ZH26	G	A	G	.	G	A	.	.	G	T	T	C	T	A	A	.	A	A	.	.	.	A	A	.
ZH29	G	A	G	A	G	A	.	.	G	T	T	C	T	A	A	.	A	A	.	.	.	A	A	.
ZH32	G	A	G	.	G	A	G	.	G	T	T	C	T	A	A	.	A	A	.	.	.	A	A	.
ZH40	G	A	G	.	G	A	.	.	G	T	T	C	T	A	A	.	A	A	.	.	.	A	A	.

nt.	370	383	391	399	401	402	403	406	409	412	424	428	430	437	441	453	457	471	492	495	496	497	507
D. simulans																							
M1	G	C	T	A	C	A	T	A	A	T	T	T	C	A	T	A	A	T	G	G	A	A	T
M4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.
M24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M37	.	.	.	G	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.
M50	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	A	.	.	.
M242	C	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.
M252	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.
M258	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.
D. melanogaster																							
ZH13,19,23A,28	.	A	A	.	A	G	G	G	G	.	A	A	A	T	.	.	T	G	A	.	C	G	G
ZH18,21A	.	A	A	.	A	G	G	G	G	.	A	A	A	T	.	.	T	G	A	.	C	G	G
ZH20A	.	A	A	.	A	G	G	G	G	.	A	A	A	T	.	.	T	G	A	.	C	G	G
ZH20B	.	A	A	.	A	G	G	G	G	.	A	A	A	T	.	.	T	G	A	.	C	G	G
ZH21B,27	.	A	A	.	A	G	G	G	G	.	A	A	A	T	.	.	T	G	A	.	C	G	G
ZH23B	.	A	A	.	A	G	G	G	G	.	A	A	A	T	C	.	T	G	A	.	C	G	G
ZH26	.	A	A	.	A	G	G	G	G	.	A	A	A	T	.	.	T	G	A	.	C	G	G
ZH29	.	A	A	.	A	G	G	G	G	.	A	A	A	T	.	.	T	G	A	.	C	G	G
ZH32	.	A	A	.	A	G	G	G	G	.	A	A	A	T	.	.	T	G	A	.	C	G	G
ZH40	.	A	A	.	A	G	G	G	G	.	A	A	A	T	.	.	T	G	A	.	C	G	G

APPENDIX B

ALIGNMENT OF SEQUENCED *D. yakuba Ran-like* ALLELES

Alignment of sequenced *D. yakuba Ran-like* alleles and *D. melanogaster Ran-like*

```
Tai-30_ran-like       ATGCAA------------GAGGTGACCTCATTCAAGCTGGTTCTTCTCGGAGACGGAGGA 48
Tai-21_ran-like       ATGCAA------------GAGGTGACCTCATTCAAGCTGGTTCTTCTCGGAGACGGAGGA 48
Tai-15_ran-like       ATGCAA------------GAGGTGACCTCATTCAAGCTGGTTCTTCTCGGAGACGGAGGA 48
Tai-159_ran-like      ATGCAA------------GAGGTGACCTCATTCAAGCTGGTTCTTCTCGGAGACGGAGGA 48
Tai-18_ran-like       ATGCAA------------GAGGTGACCTCATTCAAG-TGGTTCTTCTCGGAGACGGAGGA 47
Tai-6_ran-like        ATGCAA------------GAGGTGACCTCATTCAAG-TGGTTCTTCTCGGAGACGGAGGA 47
Tai-26_ran-like       ATGCAA------------GAGGTGACCTCATTCAAG-TGGTTCTTCTCGGAGACGGAGGA 47
Tai-27_ran-like       ATGCAA------------GAGGTGACCTCATTCAAGCTGGTTCTTCTCGGAGACGGAGGA 48
Ran-like_D.mel        ATGCAACCTCAAGAGGAAGTGAAGGCCATTTTCAAGCTGATTCTAATCGGAGACGGGGGA 60
                      ******           * *  * **   ****** ** ****  ********** ***

Tai-30_ran-like       ACTGGGAAAGCCACATTTATCAAGCGACACCTGACCGGCGAGTTCGAGAGGCGATACATT 108
Tai-21_ran-like       ACTGGGAAAGCCACATTTATCAAGCGACACCTGACCGGCGAGTTCGAGAGGCGATACATT 108
Tai-15_ran-like       ACTGGGAAAGCCACATTTATCAAGCGACACCTGACTGGCGAGTTCGATAGGCGATACATT 108
Tai-159_ran-like      ACTGGGAAAGCCACATTTATCAAGCGACACCTGACCGGCGAGTTCGAGAGGCGATACATT 108
Tai-18_ran-like       ACTGGGAAAGCCACATTTATCAAGCGACACCTGACCGGCGAGTTCGAGAGGCGATACATT 107
Tai-6_ran-like        ACTGGGAAAGCCACATTTATCAAGCGACACCTGACCGGCGAGTTCGAGAGGCGATACATT 107
Tai-26_ran-like       ACTGGGAAAGCCACATTTATCAAGCGACACCTGACCGGCGAGTTCGAGAGGCGATACATT 107
Tai-27_ran-like       ACTGGGAAAGCCACATTTATCAAGCGACACCTGACCGGCGAGTTCGAGAGGCGATACATT 108
Ran-like_D.mel        ACTGGGAAAACCACATTGGTCAAGCGACATCTGACCGGCGAGTTCAAGATGCAATACAAT 120
                      ********* *******  ********** ***** ********* * * ** ***** *

Tai-30_ran-like       GCGACCCTGGGTGTGGAGGTCCATCCAATACTCTTCCACACCAACCGAGGAGTGTACCGC 168
Tai-21_ran-like       GCGACCCTGGGTGTGGAGGTCCATCCAATACTCTTCCACACCAACCGAGGAGTGTACCGC 168
Tai-15_ran-like       GCGATCCTGGGTGTGGAGGTCCATCCAATACTCTTCCACACCAACCGAGGAGTGTACCGC 168
Tai-159_ran-like      GCGACCCTGGGTGTGGAGGTCCATCCAATACTCTTCCACACCAACCGAGGAGTGTACCGC 168
Tai-18_ran-like       GCGACCCTGGGTGTGGAGGTCCATCCAATACTCTTCCACACCAACCGAGGAGTGTACCGC 167
Tai-6_ran-like        GCGACCCTGGGTGTGGAGGTCCATCCAATACTCTTCCACACCAACCGAGGAGTGTACCGC 167
Tai-26_ran-like       GCGACCCTGGGTGTGGAGGTCCATCCAATACTCTTCCACACCAACCGAGGAGTGTACCGC 167
Tai-27_ran-like       GCGACCCTGGGTGTGGAGGTCCATCCAATACTCTTCCACACCAACCGAGGAGTGTACCGC 168
Ran-like_D.mel        GCGACCTTGGGTGTGGAGGTCGAGCAATTACTGTTTAACACCAACAGAGGAGTTTTCCGC 180
                      **** * ************** * * * **** **  ******** ******* * ****

Tai-30_ran-like       TTCTATGTGTGGGACACTGCCGGTCAGGAGAAGTTCGGTGGCCTACAAGATGGGTATTAT 228
Tai-21_ran-like       TTCTATGTGTGGGACACTGCCGGTCAGGAGAAGTTCGGTGGCCTACAAGATGGGTATTAT 228
Tai-15_ran-like       TTCTATGTGTGGGACACTGCCGGTCAGGAGAAGTTCGGTAGCCTACAAGATGGGTATTAT 228
Tai-159_ran-like      TTCTATGTGTGGGACACTGCCGGTCAGGAGAAGTTCGGTGGCCTACAAGATGGGTATTAT 228
Tai-18_ran-like       TTCTATGTGTGGGACAC--------------------------AAGATGGGTATTAT 198
Tai-6_ran-like        TTCTATGTGTGGGACAC--------------------------AAGATGGGTATTAT 198
Tai-26_ran-like       TTCTATGTGTGGGACACTGCCGGTCAGGAGAAGTTCGGTGGCCTACAAGATGGGTATTAT 227
Tai-27_ran-like       TTCTATGTGTGGGACACTGCCGGTCAGGAGAAGTTCGGT--------AGACGGGTATTAT 220
Ran-like_D.mel        ATCGATGTTTGGGACACTGCCGGTCAGGAGAGGTACGGTGGCCTGCGCGATGGGTACTTC 240
                       ** **** ********                                ** ***** *

Tai-30_ran-like       GTCCAAGGTCAATGTGCCATAATAATGTTCGACGTGAGCTCGAGAATTACCTACAAGAAT 288
Tai-21_ran-like       GTCCAAGGTCAATGTGCCATAATAATGTTCGACGTGAGCTCGAGAATTACCTACAAGAAT 288
Tai-15_ran-like       GTCCAAGGTCAATGTGCCATAATAATGTTCGACGTGAGCTCGAGAATTACCTACAAGAAT 288
Tai-159_ran-like      GTCCAAGGTCAATGTGCCATAATAATGTTCGACGTGAGCTCGAGAATTACCTACAAGAAT 288
Tai-18_ran-like       GTCCAAGGTCAATGTGCCATAATAATGTTCGACGTGAGCTCGAGAATTACCTACAAGAAT 258
Tai-6_ran-like        GTCCAAGGTCAATGTGCCATAATAATGTTCGACGTGAGCTCGAGAATTACCTACAAGAAT 258
Tai-26_ran-like       GTCCAAGGTCAATGTGCCATAAAATTGTTCGACGTGAGCTCGAGAATTACCTACAAGAAT 287
Tai-27_ran-like       GTCCAAGGTCAATGTGCCATAATAATGTTCGACGTGAGCTCGAGAATTACCTACAAGAAT 280
Ran-like_D.mel        GTCCAATCACAATGTGCCATAATAATGTTCGATGTGGCCTCGTCAAATACATATAATAAT 300
                      ******   ************* * ******* ***  ****  ** *** ** ** ***
```

Alignment of sequenced *D. yakuba Ran-like* alleles and *D. melanogaster Ran-like*. (continued)

```
Tai-30_ran-like       GTGGCACGTTGGCACCGCGACTTGGTGAGGGTATGCGGCAATATTCCGATTGTTTTGTGT 348
Tai-21_ran-like       GTGGCACGTTGGCACCGCGACTTGGTGAGGGTATGCGGCAATATTCCGATTGTTTTGTGT 348
Tai-15_ran-like       GTGGCACGTTGGCACCGCGACTTGGTGAGGGTATGCGGCAATATTCCGATTGTTTTGTGT 348
Tai-159_ran-like      GTGGCACGTTGGCACCGCGACTTGGTGAGTGTATGCGGCAATATTCCGATTGTTTTGTGT 348
Tai-18_ran-like       GTGGCACGTTGGCACCGCGACTTGGTGAGGGTATGCGGCAATATTCCGATTGTTTTGTGT 318
Tai-6_ran-like        GTGGCACGTTGGCACCGCGACTTGGTGAGGGTATGCGGCAATATTCCGATTGTTTTGTGT 318
Tai-26_ran-like       GTGGCACGTTGGCACCGCGACTTGGTGAGGGTATGCGGCAATATTCCGATTGTTTTGTGT 347
Tai-27_ran-like       GTGGCACGTTGGCACCGCGACTTGGTGAGGGTATGCGGCAATATTCCGATTGTTTTGTGT 340
Ran-like_D.mel        GTGAAAAGATGGCACCGCGACTTGGTGAGAGTATGCGGCAACATACCGATTGTCATTTGT 360
                      ***  * * ******************** ********** ** ********  * ***

Tai-30_ran-like       GGAAACAAGGTGGATATCAAGCAACGGAAGGTTAGGCCCAGGCGCTTTGACTTTCATCGT 408
Tai-21_ran-like       GGAAACAAGGTGGATATCAAGCAACGGAAGGTTAGGCCCAGGCGCTTTGACTTTCATCGT 408
Tai-15_ran-like       GGAAACAAGGTGGATATCAAGCAACGGAAGGTTAGGCCCAGGCGCTTTGACTTTCATCGT 408
Tai-159_ran-like      GGAAACAAGGTGGATATCAAGCAACGGAAGGTTAGGCCCAGGCGCTTTGACTTTCATCGT 408
Tai-18_ran-like       GGAAACAAGGTGGATATCAAGCAACGGAAGGTTAGGCCCAGGCGCTTTGACTTTCATCGT 378
Tai-6_ran-like        GGAAACAAGGTGGATATCAAGCAACGGAAGGTTAGGCCCAGGCGCTTTGACTTTCATCGT 378
Tai-26_ran-like       GGAAACAAGGTGGATATCAAGCAACGGAAGGTTAGGCCCAGGCGCTTTGACTTTCATCGT 407
Tai-27_ran-like       GGAAACAAGGTGGATATCAAGCAACGGAAGGTTAGGCCCAGGCGCTTTGACTTTCATCGT 400
Ran-like_D.mel        GGCAACAAGGTGGATATCATGCATAAAAAGACTTGGAAAAAGGGTGTTGACTTTGATCGC 420
                      ** ***************** ***    ***  * **   * * *  ******** ****

Tai-30_ran-like       AAGAAAAACCTCCACTACATTGAAATGTCCGCCAAGTCAAACTATAACATTGAGAGTCCC 468
Tai-21_ran-like       AAGAAAAACCTCCACTACATTGAAATGTCCGCCAAGTCAAACTATAACATTGAGAGTCCC 468
Tai-15_ran-like       AGGAAAAACCTCCACTACATTGAAATGTCCGCCAAGTCAAACTATAACATTGAGAGTCCC 468
Tai-159_ran-like      AAGAAAAACCTCCACTACATTGAAATGTCCGCCAAGTCAAACTATAACATTGAGAGTCCC 468
Tai-18_ran-like       AAGAAAAACCTCCACTACATTGAAATGTCCGCCAAGTCAAACTATAACATTGAGAGTCCC 438
Tai-6_ran-like        AAGAAAAACCTCCACTACATTGAAATGTCCGCCAAGTCAAACTATAACATTGAGAGTCCC 438
Tai-26_ran-like       AAGAAAAACCTCCACTACATTGAAATGTCCGCCAAGTCAAACTATAACATTGAGAGTCCC 467
Tai-27_ran-like       AAGAAAAACCTCCACTACATTGAAATGTCCGCCAAGTCAAACTATAACATTGAGAGTCCC 460
Ran-like_D.mel        AAGACAAACATTTACCTCATTGAAATGTCCGCAAAGTCGAACTATAACGTGGAGAAGCCA 480
                      * ** **** *  **  *************** ***** ********* * ****  **

Tai-30_ran-like       TTCGTCTATCTGTTGCGGAAGTTGGTTGATGATCCCAACTTGCAATTGGTCAAGAACCCC 528
Tai-21_ran-like       TTCGTCTATCTGTTGCGGAAGTTGGTTGATGATCCCAACTTGCAATTGGTCAAGAACCCC 528
Tai-15_ran-like       TTCGTCTATCTGTTGCGGAAGTTGATTGATGATCCCAACTTGCAATTGGTCAAGAACCCC 528
Tai-159_ran-like      TTCGTCTATCTGTTGCGGAAGTTGGTTGATGATCCCAACTTGCAATTGGTTAAGAACCCC 528
Tai-18_ran-like       TTCGTCTATCTGTTGCGGAAGTTGGTTGATGATCCCAACTTGCAATTGGTCAAGGACCCC 498
Tai-6_ran-like        TTCGTCTATCTGTTGCGGAAGTTGGTTGATGATCCCAACTTGCAATTGGTCAAGGACCCC 498
Tai-26_ran-like       TTCGTCTATCTGTTGCGGAAGTTGGTTGATGATCCCAACTTGCAATTGGTCAAGAACCCC 527
Tai-27_ran-like       TTCGTCTATCTGTTGCGGAAGTTGGTTGATGATCCCAACTTGCAATTGGTCAAGAACCCC 520
Ran-like_D.mel        TTCGTCTATCTATTGCGAAAATTGGTGGGTGATCCCAGCCTGCAGTTAGTCCAGAGCCCC 540
                      ********** ***** ** *** * * ******** * **** ** **  **  ****

Tai-30_ran-like       GCTCTAAAACCCCCAGAAGTGGTTTTTACCGACGA----CG------------------- 565
Tai-21_ran-like       GCTCTAAAACCCCCAGAAGTGGTTTTTACCGACGA----CG------------------- 565
Tai-15_ran-like       GCTCTAAAACCCCCAGAAGTGGTTTTTACCGACGA----CG------------------- 565
Tai-159_ran-like      GCTCTAAAACCCCCAGAAGTTGTCTTTACCGACGAGATGCGCCGTCAAGTGGAACGCGGG 588
Tai-18_ran-like       GCTCTAAAACCCCCAGAAGTGGTTTTTACCGACGAGATGCGCCGTCAAGTGGAACGCGGG 558
Tai-6_ran-like        GCTCTAAAACCCCCAGAAGTGGTTTTTACCGACGAGATGCGCCGTCAAGTGGAACGCGGG 558
Tai-26_ran-like       GCTCTAAAACTCCCAGAAGTTGTTTTTACCGACGAGATGCGCCGTCAAGTGGAACGCGGG 587
Tai-27_ran-like       GCTCTAAAACCCCCAGAAGTGGTTTTTACCGACGAGATGCGCCGTCAAGTGGAACGCGGG 580
Ran-like_D.mel        GCTATACAGCCCCCAAAAGTTGTTTTTACCGACGAGATGAGCCGTCAAGTGGAAAGCTTA 600
                      *** ** * * **** **** ** ***********     *
```

Alignment of sequenced *D. yakuba Ran-like* alleles and *D. melanogaster Ran-like*. (continued)

```
Tai-30_ran-like       --AATGGAGGCCAGCTTCTATCCTCTGCCCACTTATAACGATGATGATGATCTGTAA 620
Tai-21_ran-like       --AATGGAGGCCAGCTTCTATCCTCTGCCCACTTATAACGATGATGATGATCTGTAA 620
Tai-15_ran-like       --AATGGAGGCCAGCTTCTATCCTCTGCCCACTTATAACGATGATGATGATCTGTAA 620
Tai-159_ran-like      TTAATGGAGGCCAGCTTCTATCCTCTGCCCACTTATAACGATGATGATGATCTGTAA 645
Tai-18_ran-like       TTAATGGAGGCCAGCTTCTATCCTCTGCCCACTTATAACGATGATGATGATCTGTAA 615
Tai-6_ran-like        TTAATGGAGGCCAGCTTCTATCCTCTGCCCACTTATAACGATGATGATGATCTGTAA 615
Tai-26_ran-like       TTAATGGAGGCAAGCTTCTATCCTCTGCCCACTTATAACGATGATGATGATCTGTAA 644
Tai-27_ran-like       TTAATGGAG----------ATGTTA------------------ATGATGATGATCTGTAA 612
Ran-like_D.mel        TTCAATGAGGCCAAATCCAAACCTCTGCCCACTATTTACGATAT---TGATCTGTAA 654
                         *  ***          *   *                **     **********
```

REFERENCES

1. Adams, M. D.S. E. CelnikerR. A. HoltC. A. EvansJ. D. GocayneP. G. AmanatidesS. E. SchererP. W. LiR. A. HoskinsR. F. GalleR. A. GeorgeS. E. LewisS. RichardsM. AshburnerS. N. HendersonG. G. SuttonJ. R. WortmanM. D. YandellQ. ZhangL. X. ChenR. C. BrandonY. H. RogersR. G. BlazejM. ChampeB. D. PfeifferK. H. WanC. DoyleE. G. BaxterG. HeltC. R. NelsonG. L. GaborJ. F. AbrilA. AgbayaniH. J. AnC. Andrews-PfannkochD. BaldwinR. M. BallewA. BasuJ. BaxendaleL. BayraktarogluE. M. BeasleyK. Y. BeesonP. V. BenosB. P. BermanD. BhandariS. BolshakovD. BorkovaM. R. BotchanJ. BouckP. BroksteinP. BrottierK. C. BurtisD. A. BusamH. ButlerE. CadieuA. CenterI. ChandraJ. M. CherryS. CawleyC. DahlkeL. B. DavenportP. DaviesB. de PablosA. DelcherZ. DengA. D. MaysI. DewS. M. DietzK. DodsonL. E. DoupM. DownesS. Dugan-RochaB. C. DunkovP. DunnK. J. DurbinC. C. EvangelistaC. FerrazS. FerrieraW. FleischmannC. FoslerA. E. GabrielianN. S. GargW. M. GelbartK. GlasserA. GlodekF. GongJ. H. GorrellZ. GuP. GuanM. HarrisN. L. HarrisD. HarveyT. J. HeimanJ. R. HernandezJ. HouckD. HostinK. A. HoustonT. J. HowlandM. H. WeiC. IbegwamM. JalaliF. KalushG. H. KarpenZ. KeJ. A. KennisonK. A. KetchumB. E. KimmelC. D. KodiraC. KraftS. KravitzD. KulpZ. LaiP. LaskoY. LeiA. A. LevitskyJ. LiZ. LiY. LiangX. LinX. LiuB. MatteiT. C. McIntoshM. P. McLeodD. McPhersonG. MerkulovN. V. MilshinaC. MobarryJ. MorrisA. MoshrefiS. M. MountM. MoyB. MurphyL. MurphyD. M. MuznyD. L. NelsonD. R. NelsonK. A. NelsonK. NixonD. R. NusskernJ. M. PaclebM. PalazzoloG. S. PittmanS. PanJ. PollardV. PuriM. G. ReeseK. ReinertK. RemingtonR. D. SaundersF. ScheelerH. ShenB. C. ShueI. Siden-KiamosM. SimpsonM. P. SkupskiT. SmithE. SpierA. C. SpradlingM. StapletonR. StrongE. SunR. SvirskasC. TectorR. TurnerE. VenterA. H. WangX. WangZ. Y. WangD. A. WassarmanG. M. WeinstockJ. WeissenbachS. M. WilliamsWoodageTK. C. WorleyD. WuS. YangQ. A. YaoJ. YeR. F. YehJ. S. ZaveriM. ZhanG. ZhangQ. ZhaoL. ZhengX. H. ZhengF. N. ZhongW. ZhongX. ZhouS. ZhuX. ZhuH. O. SmithR. A. GibbsE. W. MyersG. M. Rubin, and J. C. Venter. 2000. The genome sequence of Drosophila melanogaster. Science 287:2185-2195.

2. Bai Y, Casola C, Feschotte C, Betran E. 2007. Comparative Genomics Reveals a Constant Rate of Origination and Convergent Acquisition of Functional Retrogenes in Drosophila. Genome Biology 8:R11.

3. Bayliss R, Ribbeck K, Akin D, Kent HM, Feldherr C, Gorlich D, Stewart M. 1999. Interaction between NTF2 and xFXFG-containing nucleoporins is required to mediate nuclear import of RanGDP. Jounal of Molecular Biology. 293: 579-593

4. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. 2002. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 58:899-907.

5. Betrán E, Emerson JJ, Kaessmann H, Long M. 2004. Sex Chromosomes and male functions: Where do new genes go? Cell Cycle 3:873-875.

6. Betrán E, Long M. 2003. Dntf-2r a young Drosophila retroposed gene with specific male expression under positive darwinian selection. Genetics. 164: 977-988.

7. Betrán E, Thornton K, Long M. 2002. Retroposed New Genes Out of the X in Drosophila. Genome Res. 12:1854-1859.

8. Bradley J, Baltus A, Skaletsky H, Royce-Tolland M, Dewar K, Page DC. 2004. An X-to-autosome retrogene is required for spermatogenesis in mice. Nature genetics. 36: 872-876.

9. Brosius J. 1991. Retroposons--seeds of evolution. Science. 251:753.

10. Burt A, Trivers R. 2006. Genes in conflict. The biology of selfish genetic elements. Harvard University Press, Cambridge, Massachusetts.

11. Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human diseases. Nature genetics. 39: 715-720.

12. Ciciarello M, Mangiacasale R, Lavia P. 2007. Spatial control of mitosis by the GTPase Ran. Cell Mol Life Sci 64:1891-1914.

13. Clark, A. G.M. B. EisenD. R. SmithC. M. BergmanB. OliverT. A. MarkowT. C. KaufmanM. KellisW. GelbartV. N. IyerD. A. PollardT. B. SacktonA. M. LarracuenteN. D. SinghJ. P. AbadD. N. AbtB. AdryanM. AguadeH. AkashiW. W. AndersonC. F. AquadroD. H. ArdellR. ArguelloC. G. ArtieriD. A. BarbashD. BarkerP. BarsantiP. BatterhamS. BatzoglouD. BegunA. BhutkarE. BlancoS. A. BosakR. K. BradleyA. D. BrandM. R. BrentA. N. BrooksR. H. BrownR. K. ButlinC. CaggeseB. R. CalviA. Bernardo de CarvalhoA. CaspiS. CastrezanaS. E. CelnikerJ. L. ChangC. ChappleS. ChatterjiA. ChinwallaA. CivettaS. W. CliftonJ. M. ComeronJ. C. CostelloJ. A. CoyneJ. DaubR. G. DavidA. L. DelcherK. DelehauntyC. B. DoH. EblingK. EdwardsT. EickbushJ. D. EvansA. FilipskiS. FindeissE. FreyhultL. FultonR. FultonA. C. GarciaA. GardinerD. A. GarfieldB. E. GarvinG. GibsonD. GilbertS. GnerreJ. GodfreyR. GoodV. GoteaB. GravelyA. J. GreenbergS. Griffiths-JonesS. GrossR. GuigoE. A. GustafsonW. HaertyM. W. HahnD. L. HalliganA. L. HalpernG. M. HalterM. V. HanA. HegerL. HillierA. S. HinrichsI. HolmesR. A. HoskinsM. J. HubiszD. HultmarkM. A. HuntleyD. B. JaffeS. JagadeeshanW. R. JeckJ. JohnsonC. D. JonesW. C. JordanG. H. KarpenE. KataokaP. D. KeightleyP. KheradpourE. F. KirknessL. B. KoerichK. KristiansenD. KudrnaR. J. KulathinalS. KumarR. KwokE. LanderC. H. LangleyR. LapointB. P. LazzaroS. J. LeeL. LevesqueR. LiC. F. LinM. F. LinK. Lindblad-TohA. LlopartM. LongL. LowE. LozovskyJ. LuM. LuoC. A. MachadoW. MakalowskiM. MarzoM. MatsudaL. MatzkinB. McAllisterC. S. McBrideB. McKernanK. McKernanM. Mendez-LagoP. MinxM. U. MollenhauerK. MontoothS. M. MountX. MuE. MyersB. NegreS. NewfeldR. NielsenM. A. NoorP. O'GradyL. PachterM. PapaceitM. J. ParisiM. ParisiL. PartsJ. S. PedersenG. PesoleA. M. PhillippyC. P. PontingM. PopD. PorcelliJ. R. PowellS. ProhaskaK. PruittM. PuigH. QuesnevilleK. R. RamD. RandM. D. RasmussenL. K. ReedR. ReenanA. ReilyK. A. RemingtonT. T. RiegerM. G. RitchieC. RobinY. H. RogersC. RohdeJ. RozasM. J. RubenfieldA. RuizS. RussoS. L. SalzbergA. Sanchez-GraciaD. J. SarangaH. SatoS. W. SchaefferM. C. SchatzT. SchlenkeR. SchwartzC. SegarraR. S. SinghL. SirotM. SirotaN. B. SisnerosC. D. SmithT. F. SmithJ. SpiethD. E. StageA. StarkW. StephanR. L. StrausbergS. StrempelD. SturgillG. SuttonG. G. SuttonW. TaoS. TeichmannY. N. TobariY. TomimuraJ. M. TsolasV. L. ValenteE. VenterJ. C. VenterS. VicarioF. G.

VieiraA. J. VilellaA. VillasanteB. WalenzJ. WangM. WassermanT. WattsD. WilsonR. K. WilsonR. A. WingM. F. WolfnerA. WongG. K. WongC. I. WuG. WuD. YamamotoH. P. YangS. P. YangJ. A. YorkeK. YoshidaE. ZdobnovP. ZhangY. ZhangA. V. ZiminJ. BaldwinA. AbdouelleilJ. AbdulkadirA. AbebeB. AberaJ. AbreuS. C. AcerL. AftuckA. AlexanderP. AnE. AndersonS. AndersonH. ArachiM. AzerP. BachantsangA. BarryT. BayulA. BerlinD. BessetteT. BloomJ. BlyeL. BoguslavskiyC. BonnetB. BoukhgalterI. BourzguiA. BrownP. CahillS. ChannerY. CheshatsangL. ChudaM. CitroenA. CollymoreP. CookeM. CostelloK. D'AcoR. DazaG. De HaanS. DeGrayC. DeMasoN. DhargayK. DooleyE. DooleyM. DoricentP. DorjeK. DorjeeA. DupesR. ElongJ. FalkA. FarinaS. FaroD. FergusonS. FisherC. D. FoleyA. FrankeD. FriedrichL. GadboisG. GearinC. R. GearinG. GiannoukosT. GoodeJ. GrahamE. GrandboisS. GrewalK. GyaltsenN. HafezB. HagosJ. HallC. HensonA. HollingerT. HonanM. D. HuardL. HughesB. HurhulaM. E. HusbyA. KamatB. KangaS. KashinD. KhazanovichP. KisnerK. LanceM. LaraW. LeeN. LennonF. LetendreR. LeVineA. LipovskyX. LiuJ. LiuS. LiuT. LokyitsangY. LokyitsangR. LubonjaA. LuiP. MacDonaldV. MagnisalisK. MaruC. MatthewsW. McCuskerS. McDonoughT. MehtaJ. MeldrimL. MeneusO. MihaiA. MihalevT. MihovaR. MittelmanV. MlengaA. MontmayeurL. MulrainA. NavidiJ. NaylorT. NegashT. NguyenN. NguyenR. NicolC. NorbuN. NorbuN. NovodB. O'NeillS. OsmanE. MarkiewiczO. L. OyonoC. PattiP. PhunkhangF. PierreM. PriestS. RaghuramanF. RegeR. ReyesC. RiseP. RogovK. RossE. RyanS. SettipalliT. SheaN. SherpaL. ShiD. ShihT. SparrowJ. SpauldingJ. StalkerN. Stange-ThomannS. StavropoulosC. StoneC. StraderS. TesfayeT. ThomsonY. ThoulutsangD. ThoulutsangK. TophamI. ToppingT. TsamlaH. VassilievA. VoT. WangchukT. WangdiM. WeiandJ. WilkinsonA. WilsonS. YadavG. YoungQ. YuL. ZembekD. ZhongA. ZimmerZ. ZwirkoD. B. JaffeP. AlvarezW. BrockmanJ. ButlerC. ChinS. GnerreM. GrabherrM. KleberE. Mauceli, and I. MacCallum. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203-218.

14. Clarke PR, Zhang C. 2008. Spatial and temporal coordination of mitosis by Ran GTPase. Nature reviews molecular dell biology. 9: 464-477.

15. Cushman I, Bowman BR, Sowa ME, Lichtarge O, Quiocho FA, Moore MS. 2004. Computational and biochemical identification of a nuclear pore complex binding site on the nuclear transport carrier NTF2. Journal of Molecular Biology. 344: 303-310.

16. Dorus S, Freeman ZN, Parker ER, Heath BD, Karr TL. 2008. Recent origins of sperm genes in Drosophila. MBE. 25: 2157-2166.

17. Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. Science. 303: 537-540.

18. Fay JC, Wu C. 2000. Hitchhiking under positive darwinian selection. Genetics. 155: 1405-1413.

19. Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved enconuclease required for retrotransposition. Cell. 87: 905-916.

20. Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. MBE. 11: 725-736.

21. Graur D, Li W-H. 1999. Fundamentals of Molecular Evolution. 2nd ed. Sunderland (MA): Sinauer Associates Inc.

22. Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ravi Ram K, Sirot LK, Levesque L, Artieri CG, Wolfner MF, Civetta A, Singh RS. 2007. Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. Genetics 177:1321-1335.

23. Isgro TA, Schulten K. 2007. J. Mol. Biol. 366: 330-345.

24. Jones CD, Begun DJ. 2005. Parallel evolution of chimeric fusion genes. PNAS. 102: 11373-11378.

25. Kalamegham R, Sturgill D, Siegfried E, Oliver B. 2006. Drosophila mojoless, a retroposed GSK-3, has functionally diverged to acquire an essential role in male fertility. MBE. 24: 732-742.

26. Kosakovsky SL, Frost SDW. 2005. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. MBE. 22: 478-485.

27. Kosakovsky SL, Frost SDW. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. MBE. 22: 1208-1222.

28. Kusano A, Staber C, Chan HY, Ganetzky B. 2003. Closing the (Ran)GAP on segregation distortion in Drosophila. BioEssays. 25: 108-115.

29. Kusano A, Staber C, Ganetzky B. 2002. Segregation distortion induced by wild-type RanGAP in Drosophila. PNAS. 99: 6866-6870.

30. Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nature review genetics. 4: 865-875.

31. Long M, Deutsch M, Wang W, Betran E, Brunet FG, Zhang J. 2003. Origins of new genes: evidence from experimental and computational analyses. Genetica 118:171-182.

32. Long M, Langley CH. 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science. 260: 91-95.

33. Lynch M, Conery JS. 2000. Evolutionary fate and consequences of duplicate genes. 290: 1151-1155.

34. Matsuura Y, Stewart M. 2004. Structural basis for the assembly of a nuclear export complex. Nature. 432: 872-877.

35. McCarrey JR. 1994. Evolution of tissue-specific gene expression in mammals: How a new phosphoglycerate kinase gene was formed and refined. Bioscience. 44: 20-27.

36. McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654.

37. Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

38. Okamura K, Nakai K. 2008. Retrotransposition as a source of new promoters. MBE. 25: 1231-1238.

39. Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, Eastman S, Oliver B. 2003. Paucity of genes on the Drosophila X chromosome showing male-biased expression. Science. 299: 697-700.

40. Presgraves DC. 2007. Does genetic conflict drive rapid molecular evolution of nuclear transport genes in Drosophila? Bioessays 29:386-391.

41. PresgravesDC, Stephan W. 2007. Pervasive adaptive evolution among interactors of the Drosophila hybrid inviability gene, Nup96. Mol Biol Evol 24:306-314.

42. Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. Nature reviews genetics. 3: 827-837.

43. Quimby BB, Lamitina T, L'Hernault SW, Corbett AH. 2000. The mechanism of Ran import into the nucleus by NTF2. J. Bio. Chem. 275: 28575-28582.

44. Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science 300:1742-1745.

45. Renault L, Kuhlmann J, Henkel A, Wittinghofer A. 2001. Structural basis for guanine nucleotide exchange on Ran by the regulator of chromosome condensation (RCC1). Cell 105:245-255.

46. Ribbeck K, Lipowsky G, Kent HM, Stewart M, Gorlich D. 1998. NTF-2 mediates nuclear import of Ran. EMBO Journal. 17: 6587-6598.

47. Richards S, Lounsbury K, Macara I. 1995. The C terminus of the nuclear RAN/TC4 GTPase stabilizes the GDP-bound state and mediates interactions with RCC1, RAN-GAP, and HTF9A/RANBP1. J Biol Chem 270:14405-14411.

48. Richler C, Soreq H, Wahrman J. 1992. X inactivation in mammalian testis is correlated with X-specific transcription. Nature genetics. 2: 192-195.

49. Rohozinski J, Lamb DJ, Bishop CE. 2006. UTP14c is a recently acquired retrogene associated with spermatogenesis and fertility in man. Biology of Reproduction. 74: 644-651.

50. Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497.

51. Seewald MJ, Korner C, Wittinghofer A, Vetter IR. 2002. RanGAP mediates GTP hydrolysis without an arginine finger. Nature 415:662-666.

52. Stewart M, Kent HM, McCoy AJ. 1998. Structural basis for molecular recognition between nuclear transport factor 2 (NTF2) and the GDP-bound form of the Ras-family GTPase Ran. J Mol Biol 277:635-646.

53. Tajima F. 1989. Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. Genetics. 123: 585-595.

54. Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol 21:36-44.

55. Tang S, Presgraves DC. 2009. Evolution of the *Drosophila* nuclear pore complex results in multiple hybrid incompatibilities. Science. 323: 779-782.

56. Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.

57. Vetter IR, Arndt A, Kutay U, Gorlich D, Wittinghofer A. 1999. Structural view of the Ran-Importin beta interaction at 2.3 A resolution. Cell 97:635-646.

58. Wang W, Zhang J, Alvarez C, Llopart A, Long M. 2000. The origin of the jingwei gene and the complex modular structure of its parental gene, yellow ermperor, in Drosophila melanogaster. MBE. 17: 1294-1301.

59. Wu C, Xu EY. 2003. Sexual antagonism and X inactivation - the SAXI hypothesis. TRENDS in Genetics. 19: 243-247.

60. Yang Z, Nielsen R, Goldman N, Pedersen AMK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155: 431-449.

61. Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. MBE. 17: 32-43.

62. Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13: 555-556.

63. Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. MBE. 15: 568-573.

64. Yuan X, Miller M, Belote JM. 1996. Duplicated proteasome subunit genes in Drosophila melanogaster encoding testes-specific isoforms. Genetics. 144: 147-157.

BIOGRAPHICAL INFORMATION

Charles D. Tracy was born in Dallas, Texas in 1981. He received his B.S. and M.S. degrees from the University of Texas at Arlington in 2006 and 2009 respectively. All of his degrees are in Biology with an emphasis in molecular biology and evolution. His research has focused on molecular evolution of *Drosophila* retrogenes. From 2006 to 2009, he was a laboratory instructor in the Biology Department of the University of Texas at Arlington. Future plans for Charles include laboratory research assistant and community college faculty positions. He is a member of phi sigma honors society for biology graduate students.