Convergently recruited nuclear transport retrogenes are male biased in expression and evolving under positive selection in Drosophila

Charles Tracy, Javier Río, Mansi Motiwale, Shawn M. Christensen and Esther Betrán

Department of Biology, University of Texas at Arlington

Corresponding address:

Esther Betrán

Biology Department

Box 19498

University of Texas at Arlington

Arlington, TX 76019

Phone (817) 272 1446; Fax (817) 272 2855

E-mail: betran@uta.edu

Keywords: retrogene, nuclear transport, genomic conflict, male meiotic X inactivation, sexual antagonism, dosage compensation

Running head: Convergent recruitment of nuclear transport retrogenes
The analyses of gene duplications by retroposition have revealed an excess of male-biased duplicates generated from X chromosome to autosomes in flies and mammals. Investigating these genes is of primary importance in understanding sexual dimorphism and genome evolution. In a particular instance, X-linked nuclear transport genes \((Ntf-2 \text{ and } ran)\) have given rise to autosomal retroposed copies three independent times in Drosophila (along the lineages leading to \(D. \text{ melanogaster, D. ananassae} \text{ and D. grimshawi}\)). Here we explore in further detail the expression and the mode of evolution of these Drosophila \(Ntf-2\) and \(ran\) derived retrogenes. Five out of the six retrogenes show male-biased expression. The \(ran-like\) gene of \(D. \text{ melanogaster}\) and \(D. \text{ simulans}\) has undergone recurrent positive selection. Similarly, in \(D. \text{ ananassae}\) and \(D. \text{ atripex}\), the \(Ntf-2\) and \(ran\) retrogenes show evidence of past positive selection. The data suggest that strong selection is acting on the origin and evolution of these retrogenes. Avoiding male meiotic X inactivation, increasing level of expression of X-linked genes in male testes, and/or sexual antagonism might explain the recurrent duplication of retrogenes from X to autosomes. Interestingly, the \(ran-like\) in \(D. \text{ yakuba}\) has mostly pseudogenized alleles. Disablement of the \(ran-like\) gene in \(D. \text{ yakuba}\) indicates turnover of these duplicates. We discuss the possibility that \(Dntf-2r\) and \(ran-like\) might be involved in genomic conflicts during spermatogenesis.
INTRODUCTION

Convergent duplications of two genes involved in nuclear transport, Ntf-2 and ran, in different lineages of Drosophila have been recently described (Bai et al. 2007). The Drosophila Ntf-2 (Dntf-2) appears to have generated retroposed copies—retrogene (Brosius 1991)—three independent times. The retrogenes are observed in different genomic locations in different Drosophila species with respect to the six evolutionarily conserved chromosomal arms of Drosophila (i.e., Muller elements A to F; Powell 1997): (1) Dntf-2r located in the 2L chromosomal arm (Muller element B) of the four species of the D. melanogaster complex (D. melanogaster, D. simulans, D. mauritiana and D. sechellia; Betran and Long 2003); (2) an unnamed retrogene in the D. ananassae lineage (GF23973) located in the chromosomal arm corresponding to Muller element D; and (3) a retrogene in the lineage leading to D. grimshawi (GH16820) located in the arm that corresponds to Muller element D. We call the Dntf-2 derived retrogenes of the D. ananassae and D. grimshawi lineages Da_Ntf-2r and Dg_Ntf-2r, respectively. The ran gene has also given rise to retrogenes three independent times—along some of the same lineages as Dntf-2 derived retrogenes. It gave rise to ran-like, which is present in all the species of the D. melanogaster subgroup and is located in the 3L arm (Muller element D). D. ananassae and D. grimshawi also have independently derived ran retrogenes, GF24088 and GH16204, located in Muller elements B and D, respectively. We call the ran-derived retrogenes of the D. ananassae and D. grimshawi lineages Da_ran-like and Dg_ran-like, respectively.
The parental *ran* and *Nif-2* proteins physically interact with each other and play a central role in the transport of proteins into and out of the nucleus (Ribbeck et al. 1998). Both proteins are highly conserved and are required in all eukaryotes (Quimby et al. 2000). The presence of duplicates of both genes in several lineages may have an adaptive explanation, particularly if the expression of each duplicate overlaps the other. While the function of the duplicates is not known, all six independent retrogene events occurred from X-to-autosome locations—a direction of retroduplication known to be over-represented in *Drosophila*, human, and mouse genomes (Betran et al. 2002; Emerson et al. 2004). The X-to-autosome duplications may be the result of positive selection due to pressures relating to male meiotic X inactivation, sexual antagonism, and/or X chromosome dosage compensation (see below).

In the case of male meiotic X inactivation (Hense et al. 2007; Lifschytz and Lindsley 1972), it is thought to be beneficial to have gene copies off the X and onto autosomes where transcription can occur during X chromosome inactivation (Betran et al. 2002). It has been previously shown that the *D. melanogaster* retrogene *Dntf-2r* exhibits a strong testis biased expression (Betran and Long 2003), and the same is true for *ran-like* (Chintapalli et al. 2007). In addition, both *Dntf-2r* and *ran-like* represent X-to-autosome movements (Bai et al. 2007). Thus, it is possible that the multiple parallel retropositions to autosomes of *Dntf-2* and *ran* have been driven by the selective advantage of providing a way of retaining nuclear transport function in males during meiotic X inactivation. Male meiotic X inactivation might be a consequence of the
inactivation that unsynapsed meiotic chromosomes suffer (TURNER 2007; TURNER et al. 2005).

On the other hand, under the standard sexual antagonistic hypothesis, males and females need the same gene, but alleles that are good for males might be bad for females and vice versa (i.e., some homologous traits are selected in different directions in the different sexes; CHIPPINDALE et al. 2001). In this case, if mutations are partially dominant, it might be beneficial to have the male-biased genes located on the autosomes, as opposed to the X chromosome. The X chromosome spends two-thirds of its time in females; thus, the female would be able to outcompete the male for alleles on the X chromosome (CHARLESWORTH 1987; RANZ et al. 2003; RICE 1984).

In the case of X chromosome dosage compensation, it is known that compensation in Drosophila occurs through hypertranscription of the X chromosome in males. The need for additional increased X-linked gene expression in males could be a pressure to duplicate X-linked genes onto the autosomes (VICOSO and CHARLESWORTH 2009).

Additionally, the duplicates attaining novel male testes functions could also explain the multiple duplication events of the nuclear transport genes Dntf-2 and ran. For example, it has been proposed that nuclear transport genes (including Dntf-2r) might play a role in segregation distortion and/or germline genomic conflicts involving transposable elements and viruses (PRESGRAVES 2007; PRESGRAVES and STEPHAN 2007).
Interestingly, there is direct evidence that genes involved in nuclear transport play a role in the young SD segregation distortion system in spermatogenesis of *D. melanogaster* (see review (KUSANO et al. 2003)). The main distorter, *Sd*, is a truncated duplicate form of the nuclear transport gene *RanGAP* that (mis)localizes in the nucleus (KUSANO et al. 2003). Nuclear transport genes (e.g., *RanGAP* and six different nucleoporins), and their duplicates (e.g., *Dntf-2r*), have been hypothesized to be involved in segregation distortion in Drosophila (PRESGRAVES 2007). The arms race between distorters and suppressors of distortion could lead to the fixation of duplicate genes with testes expression and to fast gene evolution of these genes (BURT and TRIVERS 2006; KUSANO et al. 2002; PRESGRAVES 2007). Indeed, *RanGAP*, six nucleoporins, and *Dntf-2r* all have been shown to have evolved under positive selection (BETRAN and LONG 2003; PRESGRAVES 2007; PRESGRAVES and STEPHAN 2007). If this is the case, the selective advantage for the fixation, longevity, and fast evolution of duplicates like *Dntf-2r* will persist as long as segregation distortion conflicts involving nuclear transport keep emerging.

Transposable elements and viruses may also impose genomic conflicts involving nuclear transport since these elements often need to enter germline nuclei to replicate (PRESGRAVES and STEPHAN 2007).

Previous work demonstrated that the *D. melanogaster Dntf-2r* has testis-biased expression and is under positive selection. *D. melanogaster ran-like* is similarly testes-biased in expression. Here we explore the mode of evolution and the pattern of expression of other independently retroposed copies of *Ntf-2* and *ran*. In both *D.*
ananassae and D. grimshawi, we show that Dntf-2 and ran retrogens are, for the most part, strongly male-biased in their transcription. In addition, we reveal that all the retrogens are evolving faster than the parental genes and are largely under positive selection. The ran-like gene is evolving under recurrent positive selection in several branches and for particular residues. In the lineages where ran-like is undergoing positive selection, some of the inferred selected amino acid changes appear to disturb specific protein-protein interaction surfaces. Polymorphism data for ran-like in the D. yakuba lineage indicate that D. yakuba’s ran-like is a pseudogene in most alleles (9 out of 10) and is wrongly annotated as functional in FlyBase. The retrogens in the D. ananassae lineage, Da_Ntf-2r and Da_ran-like, are also under positive selection. These data reveal strong selective pressures on the origin and evolution of Dntf-2 and ran retrogens.

MATERIALS AND METHODS

Strains used and DNA sequencing

Genomic DNA was extracted from single flies using a Puregene kit (Qiagen; Valencia, CA). The ran-like gene was PCR amplified from genomic DNA from 12 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 ten D. yakuba strains (Taï6, Taï15a, Taï15b, Taï18, Taï21, Taï26, Taï27, Taï30, Taï37 and Taï59) from the Taï forest in Ivory Coast. These strains were
kindly provided by the Wu, the Aquadro, the Long and the Llopart laboratories.

Oligonucleotide primers used for these reactions are given in Supplementary Table 1.

*Da_Ntf-2r* and *Da_Ran-like* were PCR amplified from genomic DNA from 10 *D. ananassae* flies from different 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* flies from different strains (14024-0361.00, 14024-0361.01, 14024-0361.02, and 14024-0361.03) obtained from the San Diego Stock Center. Oligonucleotide primers used in these reactions are given in Supplementary Table 1. *Da_Ntf-2r* sequences of *D. atripex* are partial gene sequences (357 bp long).

PCR products were sequenced from both strands using an automated DNA sequencer and fluorescent BigDye terminators (Applied Biosystems, Foster City, CA). Internal primers were used in addition to the primers above to complete some of the sequencing. Heterozygotes were cloned using a TOPO cloning kit (Invitrogen, Carlsbad, CA), and one insert was sequenced to resolve haplotypes. Accession numbers for the sequences obtained are XHHHHHHH-HHHHHHHH.

Strains that were used for the whole genome sequence project of *D. ananassae* and *D. grimshawi* (CLARK et al. 2007) were used to study pattern of expression of *Dntf-2*, *ran*, and their retrogenes in these species (see below). These strains were obtained from the former Tucson Drosophila Stock Center (currently the San Diego Stock Center).
Sequence analyses

The nucleotide sequences of the Dntf-2 gene in the 12 sequenced Drosophila species (CLARK et al. 2007) were retrieved from FlyBase along with the sequences of the paralog Dntf-2r in D. melanogaster, D. simulans, and D. sechellia. These 15 sequences were aligned with each other and with Dntf-2 retroposed sequences from D. ananassae and D. grimshawi with Clustal W software (THOMPSON et al. 1994). Sequences were analyzed initially using the CODEML software package implemented in PAML 4 (YANG 2007). The tree provided is shown in Supplementary Figure 1. Branch models were employed to determine selective pressures on the parental and retroposed sequences by calculation of $K_A/K_S$ ratios for each branch (YANG 1998). The one-ratio model sets all branches to evolve at the same rate. Additional models allowed for the genes to evolve at different rates, for the branches to evolve at different rates after duplication, or for the parental gene to evolve at a different rate after duplication. Models where $K_A/K_S$ is set at one in particular lineages were also compared to detect purifying selection. These were all a priori hypotheses that we wanted to test. These models were compared by calculating two times the log likelihood values and comparing this value to a $\chi^2$ distribution with degrees of freedom equaling the difference in number of parameters estimated by each model.

Next, site models (NSsites) of CODEML software implemented in PAML were used to uncover the possibility of positive selection acting on a few sites. Site-specific
likelihood models M7 and M8 were applied to the sequences with the appropriate tree topology (NIELSEN and YANG 1998; YANG et al. 2000). Model M7, which does not allow for sites under positive selection, was compared to model M8, which does allow for sites under positive selection. M7 assumes a beta distribution for $\omega$ between 0 and 1 over all sites, while M8 adds an additional site class ($\omega \geq 1$), with $\omega$ estimated from the data. A likelihood ratio test was performed by calculating two times the log likelihood values and comparing this value to a $\chi^2$ distribution with two degrees of freedom. Posterior probabilities of codons under positive selection were computed in model M8 using Bayes Empirical Bayes when the LRT was significant. The tree provided for the site analyses was (((Dntf-2r_D.simulans, Dntf-2r_D.sechellia), Dntf-2r_D.melanogaster).

The same PAML methods and comparisons were employed for ran and its retroposed sequences. We used parental and retrogene sequences that are present in the same species, with the addition of a ran-like ortholog in D. erecta. The ran-like ortholog in D. yakuba was not added to the analyses because a careful look at the annotation and polymorphism data revealed that it is likely a pseudogene. The tree provided for the $K_s/K_s$ ratio analyses in different branches is shown in Supplementary Figure 2. The fact that the ran-like ortholog in D. yakuba is evolving differently, similar to a pseudogene, prompted us to investigate if the other ran-like lineages evolve at different rates. The tree provided for the site-specific analyses was (((ran-like_D.simulans, ran-like_D.sechellia), ran-like_D.melanogaster), ran-like_D.erecta).
Sequences of Dntf-2r and ran-like were also analyzed using the HyPhy package. Dntf-2r sequences and tree topology were ((Dntf-2r_D.simulans, Dntf-2r_D.sechellia), Dntf-2r_D.melanogaster). The ran-like sequences and tree topology were (((ran-like_D.simulans, ran-like_D.sechellia), ran-like_D.melanogaster), ran-like_D.erecta). These sequences were uploaded to the HyPhy package available at http://www.datamonkey.org (POND and FROST 2005). Random effects likelihood (REL) and fixed effects likelihood (FEL) analyses were performed in an attempt to detect positively selected codons in the Dntf-2r and ran-like phylogenies. A Bayes factor threshold of 50, which corresponds to very low probability (P~1/Bayes factor), was used for the REL analysis, and P<0.10 for the FEL analysis (KOSAKOVSKY POND and FROST 2005). These codon analyses are presumed to be more realistic than the PAML site models because they allow for synonymous rate variation across sites (KOSAKOVSKY POND and FROST 2005).

A McDonald-Kreitman test (MCDONALD and KREITMAN 1991) was performed for the polymorphism data obtained for ran-like in D. melanogaster and D. simulans. Sequenced products were aligned using Clustal W (THOMPSON et al. 1994) and imported into DnaSP 5.0 (LIBRADO and ROZAS 2009) to perform the test. Lineage specific McDonald-Kreitman tests were also performed for ran-like in D. melanogaster and D. simulans (AKASHI 1995; PRESGRAVES 2007). A modified McDonald-Kreitman test was also used. The latter test removes synonymous unpreferred changes from the calculations because they might contribute more to polymorphisms than to divergence (AKASHI 1995; SCHLENKE and BEGUN 2003).
*D. yakuba* ran-like sequences obtained in this work were aligned with two genes of *ran-like* of *D. yakuba* annotated in FlyBase (GE22850; 3L random, and GE19852; 3L) to reveal the disablements that we detected in nine of the ten alleles and all the genome sequences (see Results and Supplementary Figure 3). We infer that the two annotated genes of *ran-like* of *D. yakuba* in FlyBase are likely alleles of the same gene. Both genes are flanked by the same region but have different disablements that likely prevented initial assembly.

Standard McDonald-Kreitman tests (McDonald and Kreitman 1991) were also performed for the polymorphism data obtained for *Da_ran-like* and *Da_Ntf-2r* in *D. ananassae* and *D. atripex*. Sequenced products were aligned using Clustal W (Thompson et al. 1994) and imported into DnaSP 5.0 (Librado and Rozas 2009) to perform the tests. The lineage specific and the modified McDonald-Kreitman tests were not performed in this case because of the absence of a close enough outgroup sequence and because of a lack of information on codon preferences in these species.

The action of recent positive selection can be addressed using proposed population genetics statistics that tests for skew in the frequency spectrum of alleles when compared to the neutral model. Tajima’s *D* (Tajima 1989) compares $\theta_r$ (the average number of nucleotide differences per site between two random sequences) and $\theta_w$ (Watterson’s estimate of $\theta$ from the number of segregating sites; Watterson 1975). Differences between $\theta_r$ and $\theta_w$ (Tajima’s *D*) reveal non-equilibrium conditions in the
history of the gene. Negative significant values are consistent with recent events of positive selection. The H statistic, the difference between $\theta_r$ and $\theta_H$ estimates (FAY and Wu 2000), measures the excess of derived variants at high frequency. Again, negative significant values of this statistic are consistent with recent events of positive selection. Tajima’s $D$ and Fay and Wu’s $H$ values were tested using neutral coalescence simulations. They were computed and tested by 10,000 simulations using DNAsp 5.1 (Librado and Rozas 2009). Recombination rates for all those simulations were considered zero to be conservative. When these tests were applied to the ran-like polymorphism data in D. yakuba, deletions were manually removed and coded as nucleotide changes.

**Expression analyses**

Tissues were homogenized, and total RNA was prepared as described by the Qiagen protocol (RNeasy kit; Qiagen, Valencia, CA). Total RNA was obtained from 30 virgin gonadectomized females, 30 gonadectomized males, 30 ovaries, and 100 testes plus accessory glands of D. ananassae. Gonadectomized males and females are flies from which we removed ovaries/testes and accessory glands by dissecting mature males and females in saline solution. After dissection, tissues were preserved in RNA-later solution (Applied Biosystems/Ambion, Austin, TX) at –20ºC after soaking them at 4ºC overnight until they were processed. Total RNA was also obtained from two virgin females and three males from D. grimshawi.
RT-PCR was conducted on total RNA for *Dntf-2*, *ran*, and the retrogenes derived from these genes in *D. ananassae* and *D. grimshawi*. Gapdh2 and Gapdh were used as internal standards for the quantitative real-time RT-PCR (qRT-PCR) performed for the retrogenes in *D. ananassae* and *D. grimshawi* respectively. Analysis of expression of intronless genes (such as *Dntf-2* and *ran* retrogenes) is challenging because genomic contamination can produce a band of the same size as that expected from the cDNA. Therefore, we digested possible contaminating DNA from the total RNA (DNase I amplification grade; Invitrogen, Carlsbad, CA) and ran controls including DNA digested total RNA without reverse transcriptase (RT-). Single strand complementary DNA (cDNA) was synthesized using Superscript and oligo-dT (Promega, Madison, WI). RT-PCRs were carried out using specific oligonucleotide primers given in Supplementary Table 1. qRT-PCRs were performed using the ABI 7300 Real Time PCR system and the SYBR Green PCR Core reagents from Applied Biosystems (Foster City, CA). Two or three replicates of RT+ and RT- were performed for every retrogene, normalizing gene and RNA extraction. The primers described above were used and produced similar amplification plots in the logarithmic scale that corresponds to similar efficiency (Schmittgen and Livak 2008). RT-PCR products were run in gels to control for any spurious amplification. Threshold cycle numbers (*C_T* values) were obtained with the default ABI software parameters. *C_T* values obtained for the retrogenes were normalized by subtracting the *C_T* value of the normalizing gene (ΔC_T). The mean difference in normalized threshold cycle number (ΔC_T) in different tissues was tested using ANOVA
analyses. Post hoc Tukey tests were also performed. Fold changes in the expression were calculated using the expression $2^{-\Delta\Delta CT}$ (SCHMITTGEN and LIVAK 2008).

RESULTS AND DISCUSSION

*Dntf-2* and *ran* retrogens are mostly male biased in expression

In *D. ananassae*, RT-PCR results reveal that both parental genes, *Dntf-2* and *ran*, are transcribed equally in both males and females across all tested tissue types (data not shown). The qRT-PCR results for the *Da_Nntf-2r* and *Da_ran-like* retrogens are shown in Figures 1A and 1B. The average of the retrogene $C_T$ difference to Gapdh2 (i.e., $\Delta C_T$) and standard error are shown for each tissue type. Small values of $\Delta C_T$ correspond to high values of transcription. An ANOVA comparing values for *Da_ran-like* expression in different tissues (Figure 1A) reveals that these values are significantly different ($F(3,8)=317.5; P<10^{-5}$). Post hoc Tukey tests show that all values are statistically significantly different from each other ($P<0.05$ in all comparisons). *Da_ran-like* is highly expressed in testes and expressed at low or very low levels in gonadectomized male bodies, gonadectomized female bodies, and ovaries. Testes expression is 60-fold higher than in other tissues.

An ANOVA comparing values for *Da_Nntf-2r* expression in different tissues (Figure 1B) shows that these values are significantly different ($F(3,7)=77.472; P<10^{-4}$).
Post hoc Tukey tests reveal that all values are statistically significantly different from each other ($P<0.05$ in all comparisons), except the comparison between male and female gonadectomized body ($P=0.9939$). $Da\_Ntf-2r$ is highly expressed in testes and expressed at much lower levels in gonadectomized male bodies, gonadectomized female bodies, and ovaries. Expression in testes is 27-fold higher than in other tissues.

In *D. grimshawi* RT-PCR results reveal again that the parental genes *Dntf-2* and *ran* are transcribed equally in males and in females (data not shown). The retrogene $Dg\_Nntf-2r$ is highly expressed in males compared to females ($F_{(1,4)}=103.96$; $P=0.0005$; Figure 1C). We do not know if this bias is due to testes expression as we were not able to obtain separate tissues for *D. grimshawi*. Unlike $Dg\_Nntf-2r$, $Dg\_ran-like$ is expressed higher in females than males ($F_{(1,4)}=43.40$; $P=0.0028$; Figure 1C).

We conclude that *Dntf-2* and *ran* retrogenes are typically male biased in expression. The retrogenes are strongly testes biased in the *D. melanogaster* lineage (BETRAN and LONG 2003; CHINTAPALLI et al. 2007) and the *D. ananassae* lineage (Figure 1). $Dg\_Nntf-2r$ is male biased in *D. grimshawi* lineage (Figure 1), although it remains unknown if $Dg\_Nntf-2r$ is testes biased.

**Sequence evolution of *Dntf-2* and its retrogenes**

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 PAML analyses for *Dntf-2* genes and *Dntf-2* retrogene branches appear in Supplementary Table 2. The free-ratio model (data not shown, $l = -2507.6789$, $p = 65$) was found to be significantly better ($P = 1.787 \times 10^{-13}$) than the one-ratio model in revealing rate differences. The two-ratio model, which estimates one $K_A/K_S$ ratio for *Dntf-2* and one ratio for the *Dntf-2* retrogene branches, was found to fit the data better than the one-ratio model ($P < 1.110 \times 10^{-16}$). A four-ratio model was implemented to allow differing rates of evolution for the three retrogenes and *Dntf-2*. This model is significantly better than the two-ratio model ($P = 8.390 \times 10^{-5}$), indicating different rates of evolution among the recurrently recruited retrogenes. The retrogene in the *D. melanogaster* complex is evolving the fastest ($K_A/K_S = 0.5311$). That said, all of the retrogene evolution rates are significantly higher than the parental rate—3-21 fold higher.

The retrogene rates can be explained either by positive selection acting on the retrogenes or by relaxation of constraint in the retrogene lineages. We have evidence of positive selection acting on *Dntf-2r* provided by the McDonald-Kreitman test (BETRAN and LONG 2003) and on *Da_Ntf-2r* (see below). The results of the *Da_Ntf-2r* McDonald-Kreitman test is consistent with a high $K_A/K_S$ ratio (0.3309) estimated in the *D. ananassae* lineage. However, no lines are available from *D. grimshawi* or close relatives to gather polymorphism data, and we have no evidence supporting selection in the *D. grimshawi* lineage at this point.

A five-ratio model is an extension of the four-ratio model. It allows a branch to evolve under a different $K_A/K_S$ ratio post-duplication. When the five-ratio model was
applied to the retrogene in the *D. melanogaster* complex, it did not fit significantly better than the simpler four-ratio model. Models that considered a different rate for the parental genes post-duplication were significantly worse at fitting the data than the four-ratio model (data not shown). Four additional models were constructed to determine if the estimated $K_A/K_S$ ratio for every gene was significantly less than one to ascertain if purifying selection is acting. All four models were significantly worse at fitting the data than the four-ratio model, indicating that all of the $K_A/K_S$ values in the phylogeny are significantly less than one and that purifying selection is acting in each of the lineages.

Although it has been pointed out that $K_S$ will saturate at large evolutionary distances (e.g., at distances larger than the *D. melanogaster* group that includes *D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta* and *D. ananassae*; *Larracuente et al*. 2008), all of our Dntf-2 retrogenses are younger than the distance represented by the *D. melanogaster* group. Dntf-2* is present only in *D. melanogaster, D. simulans*, and *D. sechellia. Da_Ntf-2r is present only in D. ananassae. Dg_Ntf-2r is present only in D. grimshawi, whose lineage is estimated to be shorter than D. ananassae lineage (*Clark et al*. 2007). However, because the estimates of Dntf-2 $K_A/K_S$ listed in Supplementary Table 2 suffer from saturation, we re-estimated the $K_A/K_S$ ratio for Dntf-2 using only the *D. melanogaster* group of species. The PAML $K_A/K_S$ ratio obtained was 0.0188. This $K_A/K_S$ value confirms that the parental genes have evolved much slower than the retrogenses.
Site models (M7 and M8, and FEL) were also fitted to the data to test for positive selection acting on particular sites of Dntf-2r (see Materials and Methods). From these results we did not infer positive selection acting on any sites of the retroposed Dntf-2 sequences. However, REL analysis detected two codons (47 [Bayes factor = 52.9636], 87 [Bayes factor = 52.6307]) that are likely under positive selection. Nucleoporins are known to interact with amino acid number 47 of the parental NTF-2 (Supplementary Data File 1).

Polymorphism data for part of Da_Ntf-2r in the D. ananassae and D. atripex lineages (Supplementary Table 3) were used to perform the McDonald-Kreitman, and Tajima’s $D$ and Wu’s $H$ tests. The results of the McDonald-Kreitman test reveal a statistically significant excess of replacement substitutions (Table 1) that is usually interpreted as recurrent positive selection acting on the protein. Tajima’s $D$ and Fay and Wu’s $H$ tests were not significant for either sample set (data not shown). The above result, along with published results for Dntf-2r (BETRAN and LONG 2003), suggest that most retroposed Dntf-2 sequences have been under strong selective pressures. We discuss the possible functional consequences of these protein changes in the Supplementary discussion file 1.

**Sequence evolution of ran and its retrogenes**

To understand the mode of evolution of ran and its retrogenes, we performed sequence analyses similar to the ones performed for Dntf-2 and its retrogenes. PAML
derived log likelihood values and maximum likelihood estimates of $K_A/K_S$ ratios for the branches in *ran* and *ran-like*, as well as the other *ran* retrogenes, are given in Supplementary Table 4. A free-ratio model (data not shown) was fitted ($l = -4410.17775, \ p = 69$) 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 (e.g., 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 that 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 than the parental genes as well (9 to 10 times faster).

To determine the mode of evolution in the *D. melanogaster* subgroup immediately after duplication of *ran*, a five-ratio model was generated. Models that consider a different rate for the parental genes after duplication were significantly worse at fitting the data (data not shown), while models that consider a different rate for the retrogenes after duplication are significantly better than the previous four-ratio model ($P = 1.8873 \times 10^{-15}$). Our five-ratio model shows a marked increase in the $K_A/K_S$ (0.7023) in *ran-like* of *D. melanogaster* subgroup. 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, even though it was still evolving roughly six times faster than the
parental gene. In fact, all ratios in the five-ratio model, including this last one, are significantly smaller than one (Supplementary Table 4). Our conclusions (i.e., that retrogene evolve much faster than parental genes) remained unchanged when we re-estimated the \( \frac{K_A}{K_S} \) ratio using only the species of the \( D. \ melanogaster \) group, where estimates should have higher accuracy (i.e., \( K_S \) should not be saturated). The \( \frac{K_A}{K_S} \) ratio had a value of 0.0065 in this calculation, confirming that retrogene evolve one to two orders of magnitude faster than the parental gene (Supplementary Table 4). Again, either relaxation of constraint or positive selection could explain the large increase in evolutionary rate amongst the \( ran-like \) lineages, although additional analyses (see below) reveal that selection is more likely.

Site models (M7 vs. M8, REL and FEL) were applied to test for positive selection acting on \( ran-like \). The M8 site model allows for positively selected sites, while M7 does not. The likelihood ratio test between site models M7 and M8 was statistically significant (\( 2\Delta l = 8.885, \text{d.f.} = 2; P = 0.0118 \)), which is indicative of positive selection acting on \( ran-like \) because it reveals that model M8 fits the data significantly better than model M7. Site model M8 estimated that 36.4% of sites in the \( ran-like \) alignment experienced positive selection (\( \frac{K_A}{K_S} = 2.52 \)). Codons that are most likely under positive selection as revealed by Bayes empirical Bayes analysis (posterior probabilities \( \geq 0.95\% \)) are shown in Table 2. Random effects likelihood (REL) analyses using the Hyphy package detected 14 sites (Table 2 and Figure 2) with a Bayes factor larger than 50 (\( P<0.02 \)) that have likely been under positive selection. The more stringent FEL analyses detected only one codon (131) as being likely under positive selection (\( P = 0.0973 \)).
Polymorphism data in *D. melanogaster*, *D. simulans*, and *D. yakuba* for *ran-like* were obtained. Supplementary Table 5 shows the polymorphism data for *D. melanogaster* and *D. simulans*. McDonald-Kreitman and modified McDonald-Kreitman tests were performed for *ran-like* using polymorphism data for *D. simulans* and *D. melanogaster* (Table 3). Both McDonald-Kreitman tests showed a significantly higher ratio of replacements per substitution, as opposed to replacements per silent polymorphism. These data support recurrent positive selection acting on *ran-like* in *D. melanogaster* and *D. simulans*. Lineage specific McDonald-Kreitman tests performed for *ran-like* in *D. melanogaster* and *ran-like* in *D. simulans* were not significant (data not shown). Likewise, Tajima’s *D* and Fay and Wu’s *H* tests were not significant, indicating that the positive selection revealed above was not recent.

*D. yakuba* polymorphism data were left out of the McDonald-Kreitman analysis because most *ran-like* alleles (nine out of ten) were found to have deletions or insertions in the coding region (Supplementary Figure 3). All of the deletions and insertions result in frame shifts and/or premature stop codons. The *ran-like* gene in *D. yakuba* is likely incorrectly annotated in FlyBase as two genes (one with an intron), but they are likely two disabled alleles (Supplementary Figure 3). There are five kinds of disabling deletions and one kind of disabling insertion in our Tai *ran-like* data set. Some of the disablements overlap with the ones observed in the FlyBase sequences. These disablements are at frequencies that vary from 10% to 30% in our sample. This observation means that no single disablement has swept through the population under directional selection in
agreement with the non-significant Tajima’s $D$ value (data not shown). We estimate the age of the alleles (SLATKIN and RANNALA 2000) to be less than $\sim 1.03 \times 10^6$ generations, assuming an effective population size of $10^6$. If, in addition, we assume two generations per year in $D. yakuba$, this estimate leads to an age of 0.515 My for the alleles present at 0.30 frequency. Since $D. santomea$ and $D. yakuba$ are estimated to have diverged $\sim 400,000$ years ago (LLOPART et al. 2002), we conclude that some of the disablements may perhaps be shared between these species.

Polymorphism data were also obtained in $D. ananassae$ and $D. atripex$ for $Da\_ran-like$ (Supplementary Table 6). The McDonald-Kreitman test was performed and revealed a significantly higher ratio of replacements per silent substitution when compared to the ratio of replacements per silent polymorphism. Fisher’s exact test was applied in this case because some cells have no observations (Table 1). This result is compatible with recurrent positive selection occurring in these lineages. However, Tajima’s $D$ and Fay and Wu’s $H$ tests were not significant for either sample set (data not shown). Taken together, the site-specific analyses and McDonald-Kreitman tests for $ran-like$ in $D. melanogaster$ and $D. simulans$, and $Da\_ran-like$ in $D. ananassae$ and $D. atripex$ strongly support the action of positive selection on $ran$ retrogenes. We discuss the possible functional consequences of these protein changes in the Supplementary Discussion File 1.

Are X to autosome retrogenes convergently duplicated because of their function in genomic conflicts in Drosophila?
A strong selective force, when acting across separate lineages, can often lead to convergent evolution. *Utp14* in mammals is an example of convergent evolution that involves the recurrent emergence and recruitment of retrogenes (Bradley et al. 2004). In this case, male meiotic X inactivation was suggested as the selective force driving the convergent recruitment of *Utp14* retrogenes.

In the present case, two interacting *Drosophila* nuclear transport genes, *Dntf-2* and *ran*, gave rise to three retrogenes each, in overlapping lineages. All these retrogenes represent X-to-autosome duplications. We further show that most of the retroposed genes have a male-biased transcription and have evolved under recurrent positive selection. We reveal that strong selection is associated with the origin and evolution of the duplicates of *Dntf-2* and *ran*. That said, *ran-like* appears to be pseudogenizing in *D. yakuba*. In further support to the X to autosome bias, we have identified another independent X to autosome retroposed copy of *Ntf-2* in *Anopheles gambiae* (data not shown).

In the Introduction, we outlined several selective hypotheses that could explain the recurrent duplication of *Ntf-2* and *ran*: male meiotic X inactivation, increasing level of expression of X-linked genes in male testes, sexual antagonism, and new function (e.g., segregation distortion or other). In particular, meiotic genomic conflicts, be they meiotic drive driven or parasite driven, are a powerful force in shaping genes and genomes as these conflicts lead to an arms race and rapid evolution (Burt and Trivers 2006). Genes (or their duplicates) involved in conflicts evolve under recurrent positive
selection (KINGAN et al. 2009) and may become fixed in the populations fairly quickly (BURT and TRIVERS 2006). These genes, particularly the duplicates, might be rendered of no use when the conflicts disappear or are dealt with in a different way. Indeed, examples exist of genes responding to strong selective pressures only to disappear after the pressures on them dissipate. Seminal fluid proteins have been observed to evolve under positive selection and have high turnover (BEGUN and LINDFORS 2005). In fact, this type of evolutionary dynamic characterizes many male-biased genes (ZHANG et al. 2007), possibly because male-biased genes are more likely to be engaged in male competition, sexual antagonism, segregation distortion, and/or parasite related conflicts more often than non-male-biased genes. Very complex situations might emerge in meiotic drive systems as observed in the case of the Winters sex ratio genes in D. simulans (KINGAN et al. 2009) where three genes of a gene family—one of them not even encoding for a protein anymore—are part of a meiotic drive system. Each of the Winters sex ratio genes are under positive selection in one population or another and are polymorphic for presence-absence of alleles.

In our case, if we hypothesize that nuclear transport is routinely involved in male germline based genomic conflicts—segregation distortion or other—then this could explain recurrent duplication of Ntf-2 and ran as well as the fast evolution of the their duplicates. We postulate that the Ntf-2 and ran parental genes are (or have been) involved in male germline genomic conflicts. Different alleles of Ntf-2 and ran in the population could act either as drivers or as suppressors of the conflict system. Alleles that increase in frequency in males, but are detrimental to females, would often end up being present at
intermediate frequencies (PATTEN and HAIG 2009; RICE 1984). Such an antagonism could promote the emergence of new genes under the model proposed by Proulx and Phillips (PROULX and PHILLIPS 2006). Duplicates of Nif-2 and ran would then help alleviate the antagonism. At the same time, factors such as the male meiotic X inactivation, increasing level of expression of X-linked genes in male testes, and/or sexual antagonism may determine the location of the duplicated genes (e.g., X to autosome), as described in the Introduction. The partial loss of function that we infer (Supplementary discussion file 1) might have occurred in some lineages (e.g., ran-like in D. melanogaster subgroup species) would be explained by a segregation distortion or genomic conflict function followed by selective pressure to differentiate from the parental gene. The complete loss of ran-like in some lineages would occur if the conflicts become resolved through an alternative mechanism in that lineage (post recruitment of ran-like) or vanish altogether. The retrogene would then become quickly pseudogenized or deleted (PETROV et al. 1996).

Finally, a knockout of Dntf-2r (D. melanogaster P-element insertion line EY05573) shows no obvious male fertility effects (unpublished), supporting a degree of dispensability of the gene. 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. Functions related to conflicts caused by transposable elements or viruses should also be tested.

ACKNOWLEDGMENTS
We want to thank Elena de la Casa-Esperón, Daven Presgraves, Dave Begun, and several reviewers for comments on this work. We also thank Mao-Lien Wu, Chung-I Wu, Tessa Bauer DuMont, Chip Aquadro, Hongzheng Dai, Manyuan Long, Ana Llopart, and the Tucson and San Diego Drosophila Stock Centers for providing stocks. Research was supported by UTA startup funds and grant GM071813 from National Institutes of Health, USA, to Esther Betrán.

REFERENCES

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 139: 1067-1076.


BRADLEY, J., A. BALTUS, H. SKALETSKY, M. ROYCE-TOLLAND, K. DEWAR et al., 2004
An X-to-autosome retrogene is required for spermatogenesis in mice. Nat Genet
36: 872-876.


CHARLESWORTH, B., COYNE, J.A., AND BARTON N.H., 1987 The relatives rates of
evolution of sex chromosomes and autosomes. The American Naturalist 130:
113-146.

CHINTAPALLI, V. R., J. WANG and J. A. DOW, 2007 Using FlyAtlas to identify better

CHIPPINDALE, A. K., J. R. GIBSON and W. R. RICE, 2001 Negative genetic correlation for
adult fitness between sexes reveals ontogenetic conflict in Drosophila. Proc Natl
Acad Sci U S A 98: 1671-1675.

CLARK, A. G., M. B. EISEN, D. R. SMITH, C. M. BERGMAN, B. OLIVER et al., 2007
Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203-
218.

EMERSON, J. J., H. KAESSMANN, E. BETRAN and M. LONG, 2004 Extensive gene traffic on

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics
155: 1405-1413.

HENSE, W., J. F. BAINES and J. PARSCH, 2007 X chromosome inactivation during


**Table 1.** McDonald-Kreitman tests for *Da_Ntf-2r* and *Da_ran-like* in the *D. ananassae* and *D. atripex* lineages.

<table>
<thead>
<tr>
<th></th>
<th><em>Da_Ntf-2r</em></th>
<th></th>
<th><em>Da_ran-like</em></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Fixed</strong></td>
<td><strong>Polymorphic</strong></td>
<td><strong>Fixed</strong></td>
<td><strong>Polymorphic</strong></td>
</tr>
<tr>
<td>Replacement</td>
<td>36</td>
<td>2</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Synonymous</td>
<td>27</td>
<td>16</td>
<td>33</td>
<td>21</td>
</tr>
</tbody>
</table>

$G_{\text{Williams correction}} = 12.991$

$P = 0.0003$

Fisher’s exact test

$P_{\text{(two tailed)}} = 0.0031$
Table 2. Comparison of likely positively selected codons identified by site-specific model M8 BEB and REL analyses in *ran-like*.

<table>
<thead>
<tr>
<th>Codon</th>
<th>M8 BEB Probability</th>
<th>REL Bayes Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>48 N</td>
<td>0.911</td>
<td>572.1</td>
</tr>
<tr>
<td>49 H</td>
<td>0.955</td>
<td>558.1</td>
</tr>
<tr>
<td>58 V</td>
<td>0.810</td>
<td>52.2</td>
</tr>
<tr>
<td>81 I</td>
<td>0.748</td>
<td>53.1</td>
</tr>
<tr>
<td>92 T</td>
<td>0.785</td>
<td>54.2</td>
</tr>
<tr>
<td>93 A</td>
<td>0.913</td>
<td>598.3</td>
</tr>
<tr>
<td>94 K</td>
<td>0.825</td>
<td>53.1</td>
</tr>
<tr>
<td>95 A</td>
<td>0.958</td>
<td>469.5</td>
</tr>
<tr>
<td>131 S</td>
<td>0.939</td>
<td>583.9</td>
</tr>
<tr>
<td>140 R</td>
<td>0.960</td>
<td>467.6</td>
</tr>
<tr>
<td>200 F</td>
<td>0.985</td>
<td>3574.9</td>
</tr>
<tr>
<td>202 D</td>
<td>0.888</td>
<td>443.9</td>
</tr>
<tr>
<td>207 Y</td>
<td>0.842</td>
<td>55.3</td>
</tr>
<tr>
<td>215 F</td>
<td>0.921</td>
<td>389.1</td>
</tr>
</tbody>
</table>

Codons that had BEB posterior probabilities \( \geq 0.95 \) or a Bayes factor \( \geq 50 \) were included in the results. Bolded values represent sites identified with statistical significance by either method. Codon number and amino acids are relative to *D. simulans* full-length protein.
Table 3. McDonald-Kreitman and modified McDonald-Kreitman tests for *ran-like* in the *D. melanogaster* and *D. simulans* lineages.

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Replacement</strong></td>
<td>47</td>
<td>21</td>
</tr>
<tr>
<td><strong>Synonymous</strong></td>
<td>12 (2)</td>
<td>14 (5)</td>
</tr>
</tbody>
</table>

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

Figure 1. qRT-PCR results for the D. ananassae and D. grimshawi retrogens (Da_ran-like, Da_Ntf-2r, Dg_ran-like and Dg_Ntf-2r) are shown. The average of the retrogene CT difference to the normalizing gene (i.e., ΔCT) and standard errors are shown for every tissue type. f denotes females and m denotes males.

Figure 2. Alignment of ran and Drosophila ran-like proteins highlighting amino acid residues of known function. Arrows point to sites that have likely been under positive selection in the retrocopies (Table 2).
Supplementary figure legends

Supplementary figure 1. Tree provided for the PAML branch analyses of \textit{Dntf-2} and retrogenes.

Supplementary figure 2. Tree provided for the PAML branch analyses of \textit{ran} and retrogenes.

Supplementary figure 3. Alignment of \textit{ran-like} disabled sequences from several strains of \textit{D. yakuba}. See Materials and Methods and Results for more details.

Supplementary discussion file legends

Supplementary discussion file 1. Functional inferences from the analyses of \textit{Dntf-2} and \textit{ran} retrogene protein sequences

Supplementary data file legends

Supplementary data file 1. Alignments of \textit{Ntf-2} and \textit{ran} proteins and Drosophila retroduplicates highlighting amino acid residues of known function.

Supplementary table legends
Supplementary table 1. Oligonucleotide primers used in this work.

Supplementary table 2. PAML comparisons for Dntf-2 and Dntf-2 retrogene.

Supplementary table 3. Summary polymorphism data for Da_Ntf-2r in D. ananassae and D. atripex.

Supplementary table 4. PAML comparisons for ran and ran retrogene.

Supplementary table 5. Summary polymorphism data for ran-like in D. melanogaster and D. simulans

A. $Da\_Ran\text{-}like$

<table>
<thead>
<tr>
<th>ΔC_T</th>
<th>f-gonads</th>
<th>ovaries</th>
<th>m-gonads</th>
<th>testes</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B. $Da\_Ntf\text{-}2r$

<table>
<thead>
<tr>
<th>ΔC_T</th>
<th>f-gonads</th>
<th>ovaries</th>
<th>m-gonads</th>
<th>testes</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C. $ΔC_T$

<table>
<thead>
<tr>
<th>ΔC_T</th>
<th>Dg_Ntf-2r_f</th>
<th>Dg_Ntf-2r_m</th>
<th>Dg_Ran-like_f</th>
<th>Dg_Ran-like_m</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>