Intended as Discovery Article

Relocation Facilitates the Acquisition of Short Cis-Regulatory Regions that Drive the Expression of Retrogenes during Spermatogenesis in Drosophila

Mehran Sorourian[#], Mansi M. Kunte[#], Susana Domingues, Miguel Gallach[§],

Fulya Özdil[¶], Javier Río, 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, USA

Phone (817) 272 1446; Fax (817) 272 2855; E-mail: betran@uta.edu

[#] These authors contributed equally to this work

§ Other affiliations:

Miguel Gallach: Center for Integrative Bioinformatics Vienna (CIBIV), Max F Perutz

Laboratories. University of Vienna and Medical University of Vienna.

¶Current affiliation:

Fulya Özdil: Namık Kemal University, Faculty of Agriculture, Department of

Agricultural Biotechnology, Tekirdağ, Turkey.

Key words: Drosophila, retrogene, testis expression, regulatory element, gene

duplication, gene relocation

Running head: Short testis regulatory regions in Drosophila retrogenes

Abbreviations: TSS, transcription start site; β2-UE1, *β2-tubulin* gene upstream element 1.

ABSTRACT

Retrogenes are functional processed copies of genes that originate via the retrotranscription of a mRNA intermediate and often exhibit testis-specific expression. Although this expression pattern appears to be favored by selection, the origin of such expression bias remains unexplained. Here, we study the regulation of two young testis-specific Drosophila retrogenes, *Dntf-2r* and *Pros28.1A*, using genetic transformation and the EGFP reporter gene in *D. melanogaster*. We show that two different short (<24 bp) regions upstream of the transcription start sites (TSSs) act as testis-specific regulatory motifs in these genes. The *Dntf-2r* regulatory region is similar to the known *β2 tubulin* 14 bp testis motif (β2-UE1). Comparative sequence analyses reveal that this motif was already present before the *Dntf-2r* insertion and was likely driving the transcription of a non-coding RNA. We also show that the β2-UE1 occurs in the regulatory regions of other testis-specific retrogenes, and is functional in either orientation. In contrast, the *Pros28.1A* testes regulatory region in *D. melanogaster* appears to be novel. Only *Pros28.1B*, an older paralog of the *Pros28.1* gene family, seems to carry a similar regulatory sequence. It is unclear how the *Pros28.1A* regulatory region was acquired in *D. melanogaster*, but it might have evolved *de novo* from within a region that may have been pre-primed for testes expression. We conclude that relocation is critical for the evolutionary origin of male germline-specific cis-regulatory regions of retrogenes because expression depends on either the site of the retrogene insertion or on sequence changes close to the TSS thereafter. As a consequence we infer

that positive selection will play a role in the evolution of these regulatory regions and

can often act from the moment of the retrocopy insertion.

INTRODUCTION

Retrogenes are functional processed copies of genes (Brosius 1999) which originate when an mRNA is reverse transcribed and reinserted into the genome. At the time of insertion retrogenes are devoid of introns, have a poly-A tail and are flanked by short direct repeats. However, the last two features are quickly lost and not found in old *Drosophila* retrogenes (Betran, et al. 2002). Although the chromosome of origin of the polymorphic retrocopies appears to fit random expectations in *Drosophila* (Schrider, et al. 2011), flies exhibit an excess of retrogenes that have originated from genes on the X chromosome and are inserted in autosomes (Betran, et al. 2002; Dai, et al. 2006; Bai, et al. 2007; Meisel, et al. 2009; Vibranovski, Zhang, et al. 2009; Han and Hahn 2012). In addition, most retrogenes exhibit testis-specific expression while parental genes are broadly expressed (Betran, et al. 2002; Bai, et al. 2007; Vibranovski, Lopes, et al. 2009; Han and Hahn 2012). Similar patterns have also been observed in placental mammal and marsupial genomes (Emerson, et al. 2004; Potrzebowski, et al. 2008). The insertion and the expression patterns do not correspond to the retropositional biases across the genome and have, therefore, been attributed to the action of selection (Emerson, et al. 2004; Vinckenbosch, et al. 2006; Schrider, et al. 2011; Schrider, et al. 2013).

At the time of retroposition, all retroposed copies lack the regulatory regions (enhancer and promoter regions) of their parental genes. Thus, to be preserved in the genome, most retrocopies that will eventually become functional genes will either be inserted downstream of an existing regulatory region or quickly evolve one *de novo*.

5

It is often assumed that the insertion of a retrocopy between an existing gene and its regulatory region would perturb the gene's regulation and be deleterious. However, in the case of an insertion into a duplicated region where there might be redundancy, the insertion has a higher chance of evolving into a new gene (Long and Langley 1993; Wang, et al. 2000). The insertion site may exert particular effects on the transcription of the newly inserted gene. Initial transcription may occur if the insertion is close to transcribed genes or, in the case of testis-expressed retrogenes, in regions that are enriched with testis-expressed genes, i.e., testis neighborhoods (Kalmykova, et al. 2005; Vinckenbosch, et al. 2006; Bai, et al. 2008; Dorus, et al. 2008). Either transcriptional facilitation or co-regulation has been proposed as mechanisms to explain this effect (Loppin, et al. 2005; Vinckenbosch, et al. 2006). Transcriptional facilitation could occur if the retrocopy is inserted in an actively transcribed open chromatin region. This mechanism would increase the probability of starting transcription of the new gene using a cryptic promoter. On the other hand, gene co-regulation could take place through the sharing of enhancers (Lemos, et al. 2008) that can act long or short range and affect more than one gene (Loppin, et al. 2005; Vinckenbosch, et al. 2006). However, in this case, the evolution of a core promoter region in the retrocopy might be needed for the proper start of transcription.

In this study, we experimentally characterize the regulatory regions of two relatively young retrogenes, *Dntf-2r* (*CG10174*) and *Pros28.1A* (*CG17268*), and their evolutionary origin. We show that a short (i.e., 14 bp) cis-regulatory region similar to the β2-UE1 of the *β2-tubulin* gene (*CG9359*; (Michiels, et al. 1989) is needed for the

testis-specific expression of *Dntf-2r* in *D. melanogaster*. We also determine that the motif was present before *Dntf-2r* insertion, it is found upstream of other testis-specific retrogenes and either orientation (+ or -) can drive testis-specific expression. A novel regulatory region that includes 23 bp upstream of the TSS is needed to drive testis-specific expression of *Pros28.1A* in *D. melanogaster*. Only *Pros28.1B*, another testis-specific gene of the same gene family, seems to share part of this region. *Pros28.1B* is an older DNA mediated duplication of *Pros28.1* and it is, at this point, unclear how its regulatory region was acquired. These case studies illustrate that the evolutionary origin of male germline-specific cis-regulatory regions depends on relocating a coding region to a new context where regulatory regions for testis expression either already exist or can evolve close to the TSS.

RESULTS

***Dntf-2r* is expressed from a β2-UE1-like element**

Dntf-2r originated less than 12 Mya (Betran and Long 2003) and is present in four species of Drosophila: *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. mauritiana* (Betran and Long 2003). It shows testis- or male-biased expression in all species where it is present (Betran and Long 2003; Chintapalli, et al. 2007; Supplementary Results). To identify the regulatory region that drives the testis-specific expression of *Dntf-2r* in *D. melanogaster*, clones carrying the region upstream of the previously described TSS (Betran and Long 2003), its 5'UTR and *Dntf-2r* coding region fused to EGFP (Enhanced Green Fluorescent Protein) as a

reporter gene were transformed in *Drosophila* (see Materials and Methods). The longest construct contained 151 bp upstream of the TSS including 25 bp of the 5'UTR of the gene *bicoid stability factor* (*bsf*) located upstream of *Dntf-2r* in a head-to-head orientation (Figure 1A). The transformed flies expressed the green fluorescent fusion protein (Figure 1A and Supplementary Fig. 4A) in a pattern that mimics the Dntf-*2r* RNA in situ profile (Figure 1B) suggesting that this construct harbors the complete testis-specific regulatory region of *Dntf-2r*. Subsequently, we generated constructs carrying -101 bp, -77 bp, -50 bp, and -4 bp upstream of the TSS. Of those, only the -101 bp and -77 bp constructs show testis-specific green fluorescence (Figure 1A and Supplementary Fig. 4A-C). The others did not show fluorescence (i.e., the observed fluorescence was comparable to the autofluorescence of the white mutant control, w[1118]; Supplementary Fig. 4D-E). Independent insertion lines show similar results for each construct (Supplementary Fig. 4A-E) indicating that the constructs themselves, and not the region where the constructs are inserted, are responsible for the observations. Green fluorescence is also observed in the male gonads of the second instar larvae of the -151 bp lines (Supplementary Fig. 4F). No consistent fluorescence is observed in other tissues like the gut and ovaries for this line (Supplementary Fig. 4G) when compared to the control (w1118), thus confirming the testis-specific expression of the construct. We conclude that a 27-bp region (i.e., the region between -77 bp and -50 bp shaded in yellow in Figure 2A) is needed to drive the expression of the *Dntf-2r*-EGFP fusion gene in testis.

We observed differences in the intensity of the green fluorescence between -151 and -101 and -77, which we could not properly quantify. Specifically, flies containing the -151 bp construct appear to fluoresce stronger in testis than the flies carrying the shorter fluorescing constructs -101 bp and -77 bp, but in the same pattern. Interestingly, we found two potential BEAF-32 insulator binding sites at -135–-129 bp and -115–-109 bp. These two BEAF-32 insulator sequences appear to be used at least during embryogenesis (FlyBase and (Negre, et al. 2010) and, consistent with the previously described cases, might prevent co-expression of head-to-head genes (Yang, et al. 2012). These insulator sequences are absent in the shorter constructs and it is unclear if or how the lack of these sequences can actually affect the levels of expression of the genes and our constructs in testis. Therefore, we concluded that a 27-bp region (i.e., the region between -77 bp and -50 bp shaded in yellow in Figure 2A) is needed to drive the expression of the *Dntf-2r*-EGFP fusion gene in testis, although other upstream regions may also influence the expression level.

The 27-bp region contains a β2-UE1-like sequence and the upstream region of *Dntf-2r* also harbors a sequence (Figure 2A) that is identical to the 7-bp motif with quantitative effects described in the testis-specific *β2 tubulin* gene (GGATATT; (Michiels, et al. 1989). Those motifs are found in *D. simulans*, *D. sechellia* and *D. mauritiana* but have positional difference compared to *D. melanogaster* (Supplementary Fig. 3). We infer from the sequence conservation that these elements are most likely functional in the other species as well.

The β2-UE1-like element was present before the insertion of *Dntf-2r*

To address the evolutionary origin of the β2-UE1-like element of *Dntf-2r*, we compared the region before and after the insertion of the retrogene and attempted to reconstruct the ancestral state of this genomic location using maximum parsimony. The multiple alignment of the orthologous sequences revealed that the motifs upstream of *Dntf-2r* are highly conserved even in species that do not have the retrogene (Figure 2B). We inferred with confidence that the motif with quantitative effects (GGATATT) was present at the time of insertion and that the β2-UE1-like region was at least partially present (Figure 2B).

Indirect evidence supports that this region was functional before the insertion of *Dntf-2r* drove testis expression. First, we find that the β2-UE1-like and the motif with quantitative effects are partly and completely conserved, respectively, in *D. yakuba, D. teissieri* and *D. erecta*, where *Dntf-2r* is absent (Figure 2). In addition, the BEAF-32 binding sites are also conserved in those species (Data not shown). This supports a head-to-head arrangement of *bsf* with another gene even before the insertion of *Dntf-2r*. Recently, a non-coding RNA gene (*let-7-C*; *CR43344*) has been annotated in *D. melanogaster* downstream of *Dntf-2r*. This non-coding RNA shows a male-biased pattern of expression similar to *Dntf-2r* according to modENCODE data (Marygold, et al. 2013) and its location is conserved in *D. yakuba, D. teissieri* and *D. erecta* (Data not shown), indicating that the regulatory regions of *Dntf-2r* likely predated its origin.

To verify that this region is functional in species where *Dntf-2r* does not exist (i.e., species most likely carrying the ancestral configuration of this genomic region), we fused and transformed a construct containing the 14 bp orthologous region (see green highlight in Figure 2B) from *D. yakuba* and the rest of the region from *D. melanogaster* in front of EGFP and checked if it drives testis-specific expression (See Materials and Methods). As proposed above, we observed testis-specific fluorescence in most of the independent transformants containing the construct (Supplementary Fig. 5) and concluded that the region is capable of driving testis-specific expression at a discernible level (Figure 1C) when compared to constructs lacking the *D. yakuba* β2-UE1-like region (Supplementary Fig. 4D-E). All the evidence indicates that the region contained a testis-specific regulatory element before the insertion of *Dntf-2r*.

Pros28.1A is expressed from a novel testis-specific cis-regulatory element

Pros28.1A is estimated to be ~39 My old (Sorourian and Betran 2010). Previous work (Sorourian and Betran 2010) revealed that transgenic flies carrying 246 bp upstream of the TSS, the 5'UTR of *Pros28.1A*, and EGFP as a reporter gene show a pattern of expression that recapitulates the known testis expression pattern of the gene. Here, in an effort to elucidate in more detail the testis-specific regulatory region of this gene, 5 additional constructs were made. The constructs contained -189 bp, -80 bp, -46 bp, -23 bp and 0 bp upstream of the *Pros28.1A* TSS (Figure 3A). Only the flies transformed with the construct lacking any upstream region lacked EGFP expression in the testis (Figure 3 and Supp. Fig 6). So, we used site directed

mutagenesis to completely change the 23-bp region immediately upstream of

Pros28.1A TSS of *D. melanogaster* to make another construct (See Supplementary

Materials and Methods and Figure 3A). This construct carrying a different 23 bp

region lacked EGFP expression in the testis (Figure 3 and Supp. Fig 6). The

transformants showing the testis-specific expression pattern in adult testis showed

expression in larval male gonads (Supplementary Fig 6A-D and 6G), but not in other

tissues (Supplementary Fig. 6H). In addition, we observe a lower level of expression

as the constructs are shortened (Figure 3A and Supplementary Fig 6A-D). However,

we have not quantified this effect. The decrease in expression level could potentially

be explained by the presence of motif/s with quantitative effects upstream of the

testis-specific element as proposed for the *β2 tubulin* gene (Michiels, et al. 1989) or

the loss of unknown insulator elements in shorter constructs as we have inferred for

Dntf-2r in this work (see above). We concluded that information contained within the

23 bp upstream of *Pros28.1A* TSS (i.e., 5'GAAAAATTCATATTGTTTCAAGT3';

Figure 3B) is needed for testis-specific expression of *Pros28.1A* in *D. melanogaster*.

The 23 bp upstream of *Pros28.1A* TSS included in the shortest fluorescent

construct (Figure 3B; bold highlight) does not appear to be entirely conserved among

species that carry this retrogene but there are at least 5 bp (i.e., GAAAA) conserved.

In addition, we could not find any of the promoter elements previously described in

this region, including the *β2-tubulin* testis-specific promoter (Michiels, et al. 1989;

Ohler, et al. 2002; Fitzgerald, et al. 2006). Thus, we infer that the 23 bp region

contains a novel testis-specific cis-regulatory element.

Interestingly, there are 13 bp of this 23 bp region that show the highest similarity between *Pros28.1A* and *Pros28.1B* upstream region (i.e., 5'CATATTGTTTCAA3' *Pros28.1A* vs. 5'CATATTGTTTTAA3' *Pros28.1B*; (Sorourian and Betran 2010). The expression of both genes (*Pros28.1A* and *Pros28.1B*) is testis specific and sensitive to mutations in meiotic arrest genes (see below and (Mikhaylova, et al. 2006), and this shared sequence might potentially be responsible for the similar regulation of these two genes. However, as with *Pros28.1A*, the upstream region of *Pros28.1B* does not appear to be more conserved among species than other adjacent non-coding regions (Data not shown). This may indicate that the 23bp regulatory region of *Pros28.1B* has changed in other lineages or recently originated in the common ancestor of *D. melanogaster, D. simulans, D. sechellia* and *D. mauriatiana*. It remains possible that the few bp that are conserved between all species that have *Pros28.1A* (Figure 3B) are part of the motif/s needed for testis-specific expression in *D. melanogaster* and that the *D. melanogaster* regulatory region evolved from within a region that was pre-primed for testes expression.

The β2-UE1 motif occurs in the upstream region of other testis-specific retrogenes

We studied whether the 14-bp β2-UE1 motif of the *β2 tubulin* gene (5'ATCGTAGTAGCCTA3') and the 13-bp conserved portion between *Pros28.1A* and *Pros28.1B* (5'CATATTGTTT(C/T)AA3'), also occur in upstream regions of other genes. A preliminary BLAST analysis of both motifs found only the 14-bp β2-

UE1 motif in several testis-specific genes. We generated a position-specific scoring matrix (PSSM; Supplementary Fig. 7 and Supplementary Table 1) from the BLAST hits alignment, which we gave to GLAM2SCAN as an input matrix to scan the upstream region of all *D. melanogaster* genes (see Materials and Methods for more details). We found that 120 upstream regions of the 1,245 testis-specific genes (10%) and 476 upstream regions 13,253 non-testis-specific genes (i.e., 4%) contained a copy of the β2-UE1 motif indicating that testis-specific genes are significantly enriched for this motif ($P=2.91\times10^{-19}$; Fisher's exact test). Remarkably, 60 of the 476 non-testis-specific genes containing a copy of the β2-UE1 motif upstream of their TSS are actually testis-biased (i.e., genes that are non-testis-specific under our definition [see Materials and Methods], but whose level of transcription in testis was 10-fold higher than the average in the whole fly). Out of the 120 testis-specific genes, 13 are known testis-specific retrogenes (Betran, et al. 2002; Dai, et al. 2006; Bai, et al. 2007).

Interestingly, 54 β2-UE1-like motifs follow the same strand orientations as the testis-specific gene, yet there are 61 in the opposite strand. This observation prompted us to generate a construct that carries the upstream region and 5'UTR of the β2-tubulin gene but where the 14-bp β2-UE1 motif has been flipped (i.e., inverted; Figure 1D). Ten independent transformants were examined and showed fluorescence specifically in testis (Supplementary Fig. 8) demonstrating that the β2-UE1 motif works in both orientations and could potentially work bidirectionally.

Transcription of *Dntf-2r* and *Pros28.1A* is sensitive to mutations in meiotic arrest genes

It has been suggested that the β2-UE1-like promoter of the testis-biased gene *Sdic* is bound by the regulatory protein Modulo and it has been shown that transcription of *Sdic* is sensitive to mutations in meiotic arrest genes (e.g., *rye*) and *modulo* (*mod*; (Mikhaylova, et al. 2006). The *β2 tubulin* gene transcription is also affected by mutations in meiotic arrest genes (Mikhaylova, et al. 2006). We analyzed the transcription of *Dntf-2r* in wild-type (Oregon-R), *rye* (a meiotic arrest gene), and *mod* mutant strains by QRT-PCR (see Materials and Methods). Significant differences ($F_{(2,6)}=5.6551$; $P=0.0416$) were observed among the mean difference for the normalized threshold cycle number (ΔC_T; Supplementary Fig. 9) between strains. Post-hoc Tukey tests revealed a significant difference between Oregon-R and *rye* mutant strain ($P=0.0464$) and the same trend, although not significant ($P=0.0867$), was observed after *a posteriori* testing between Oregon-R and *mod* mutant strains. Fold-down changes in the expression were calculated using the expression $2^{-\Delta\Delta C_T}$ (Schmittgen and Livak 2008). *Dntf-2r* fold changes are similar in both mutant comparisons with Oregon-R: 1.8109 for *mod* and 2.0280 for *rye*. Although many genes can be affected in meiotic arrest mutants because spermatogenesis does not occur (White-Cooper 2010), our observation is consistent with what has been observed for genes that harbor the β2-UE1-like promoter.

We also analyzed the transcription of *Pros28.1A* in a wild-type strain (Oregon-R), and *rye* and *mod* mutant strains by QRT-PCR (see Materials and Methods). Significant differences ($F_{(2,6)}110127$; $P=0.0098$) were observed among the

mean difference for the normalized threshold cycle number (ΔC_T; Supplementary Fig. 9) between strains. Post-hoc Tukey tests revealed a significant difference between Oregon-R and the *rye* mutant strain (P=0.0181), and between the *mod* and *rye* mutants (P=0.0138). No differences were observed between Oregon-R and the *mod* mutant strain (P=0.9674). Fold-down regulatory changes in the expression were similar in both comparisons with the *rye* mutant strain: 5.2659 times lower expression compared to Oregon-R and 5.8411 times lower compared to *mod* mutant strain. We conclude that *Pros28.1A* is a testis-biased gene whose expression is sensitive to mutations in meiotic arrest genes, but not to mutations in the protein modulo. This is similar to the sensitivity observed for *Pros28.1B* transcription (a different *Pros28*.1 duplicate that also has spermatogenesis-specific expression; (Yuan, et al. 1996). *Pros28.1B* transcription was affected in the *rye* mutant strain, but only slightly affected in *mod* mutants (Mikhaylova, et al. 2006). Again, although many genes can be affected in meiotic arrest mutants because spermatogenesis does not occur (White-Cooper 2010), our observation is consistent with a similar regulatory mechanism used for these two *Pros28.1* duplicates as the effects that we tested are consistent between these genes.

DISCUSSION

Nature and origin of regulatory regions of retrogenes

Retrogenes lack regulatory regions at the time of insertion and only retrogenes that either insert downstream of an existing cis-regulatory region and promoter, or

quickly evolve a regulatory region *de novo* will be retained over the course of evolution. In this study, we report one example for each of these evolutionary scenarios.

Dntf-2r is transcribed from a preexisting testis regulatory region upstream from the insertion site. The regulatory region has 64% identity with the β2UE1 promoter of the *β2-tubulin* gene in *D. melanogaster* and also contains the 7-bp motif with quantitative effects of the *β2-tubulin* gene (Michiels, et al. 1989). These motifs were most likely present before the insertion of *Dntf-2r* and functional in all the species where *Dntf-2r* is present. We hypothesize that the β2UE1 promoter was initially driving testis-specific transcription of a non-coding RNA (*let-7-C*), still present downstream of *Dntf-2r*. Because the regulatory region of *Dntf-2r* does not seem to have duplicated, it is difficult to explain why the insertion of *Dntf-2r* does not affect the regulation of *let-7-C*. Importantly though, *Dntf-2r* insertion would have likely been expressed in testis upon insertion and, therefore, selection would have potentially acted from the very beginning to increase the frequency of this retrogene in the population until all individuals would have carried this gene.

On the other hand, the lack of sizeable conservation (only 5 bp seem to be conserved in all species that contain the gene) of the regulatory sequence of *Pros28.1.A* in different species is in agreement with the observation that TSS of *Pros28.1A* varies among species despite the conserved testis-specific expression pattern (Sorourian and Betran 2010). The order of events that eventually led to testis-

expression upon insertion of *Pros28.1A* is unclear. Our data support the current existence of a new motif in *D. melanogaster* and, potentially, in some close related species that carry this testis-specific gene but not in others in favor of a constantly evolving testis-specific regulatory region.

Consistent with the results of various transgenic experiments for many testis-specific genes (See (White-Cooper 2010) for a review), we find that very short cis-regulatory regions are required for the expression of *Dntf-2r* and *Pros28.1A* in male germline but also infer the existence of boundary elements and motif with quantitative effects around these regions. In addition, we reveal that relocation is critical for the evolutionary origin of male germline-specific cis-regulation of retrogenes because this pattern of expression depends on losing the initially broad/housekeeping expression of the parental genes (Bai, et al. 2007) and gaining short and testis-specific expression from either the site of retrogene insertion or from sequence changes close to the TSS thereafter.

The observed effect of mutations in meiotic arrest genes in the expression of *Dntf-2r* and *Pros28.1A* is consistent with the effect found in other genes carrying similar regulatory motifs (Mikhaylova, et al. 2006) and with a newly proposed regulatory complex involving *Ntx1* in Drosophila that couples nuclear export of intronless genes with the testis meiotic arrest complex and regulates the transcription of those genes specifically during spermatogenesis (Caporilli, et al. 2013).

Is there low specificity and/or high turnover of regulatory regions of testis-specific genes?

We observe that the known β2UE1-like regulatory sequences used by several testis-specific genes in *D. melanogaster* (Figure 2A) actually have low identity to each other even when gaps are allowed in the alignment. We also found that the regulatory region of *Pros28.1A* shows low conservation, most likely revealing a turnover of the regulatory regions in Drosophila testis genes. This reveals low specificity of the transcription factors binding those sites or the presence of related, but divergent, regulatory complexes and high turnover of the regulatory regions.

Why should there be low specificity and/or high turnover of regulatory regions of testis-specific genes? It is known that some of the general transcription factors (i.e., TATA-binding protein associated factors also known as TBP-associated factors or TAFs) have testis-specific isoforms (i.e., tTAFs; (Sassone-Corsi 2002; Hiller, et al. 2004) that contribute to the regulated deployment of testis-specific genes. Recently, it has been observed that the testis-specific TAFs duplicates and *mod* evolve under positive selection (Haerty, et al. 2007; Li, et al. 2009), and that a new testis-specific copy of a transcriptional cofactor has originated through duplication modifying Drosophila testis expression (i.e., a duplicate of the ortholog of human *Rcd-1*; (Quezada-Diaz, et al. 2010; Chen, et al. 2012). While the high turnover of regulatory regions (i.e., same regulation despite the lack of sequence conservation) compared to protein-coding regions might be a general feature of regulatory regions (Ludwig, et al. 2000; Hare, et al. 2008), rapid evolution under positive selection of

the testis-specific TAFs and *mod* and new regulatory gene acquisitions reveal that there are particular selective pressures in this tissue that likely drive an even faster evolutionary turnover of regulatory regions. Many genes have been observed to evolve under positive selection in testis probably as a response to male-male competition, sexual antagonism, segregation distortion, and/or parasite-related conflicts (Zhang, et al. 2004; Haerty, et al. 2007; Presgraves 2007; Presgraves and Stephan 2007). These selective pressures might also act on regulatory regions because the gain or loss of testis expression (i.e., gain or loss of a cis-regulatory region) of newly duplicated genes (Kaessmann 2010) or of antisense RNAs to regulate parasites in testis (Lankenau, et al. 1994) will be under strong selection. We predict that the evolutionary pattern observed in Drosophila male germline gene regulation should hold in other lineages.

MATERIALS AND METHODS

Drosophila stocks and fly handling

Several wild-type strains were used to PCR amplify the regulatory region of *Dntf-2r* in different species. *D. simulans* from Florida (provided by J. Coyne), *D. sechellia* (provided by J. Coyne)*, D. mauritiana* (163.1; (Lemeunier and Ashburner 1976), *D. teisseri* (118.2; (Lemeunier and Ashburner 1976), *D. yakuba* (115; (Lemeunier and Ashburner 1976), *D. erecta* (154.1; (Lemeunier and Ashburner 1976) and *D. eugracilis* (14026-0451-07; UC San Diego Drosophila Stock Center).

QRT-PCR

Tissues were homogenized and total RNA was prepared from whole males of three strains ($Taf12L^{KG00946}$ [rye], and mod^{07570} mutant strains and Oregon R) for comparison. The QIAGEN RNeasy mini kit (QIAGEN, Valencia, CA) and protocol was used for RNA extractions. Quantitative real-time RT-PCR (QRT-PCR) was conducted for the retrogenes using oligoprimers

5´AGCCACGAAGAGGGATCCTC3´and 5'TTGTCCAGCAGTACTACGCC3' for *Dntf-2r* and 5'AACTGGGTTACCTGCAGAAGG3' and

5'TGCCTGCTAACTAACCCAAAC3´ for *Pros18.1A*. *Gapdh2* was used as a normalizer in the QRT-PCR. The oligoprimers used were

5'TCAGCCATCACAGTCGATTC3' and 5'CAAACGAACATGGGAGCATC3'. QRT-PCRs were performed using the ABI 7300 Real Time PCR system and the SYBR Green PCR Core reagents from Applied Biosystems (Applied Biosystems, Foster City, CA). Three replicates of RT+ and RT- were performed for every retrogene, normalizing gene, and RNA extraction. The primers described above produced similar amplification plots in the logarithmic scale that correspond to similar efficiency (Schmittgen and Livak 2008). QRT-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 for the normalized threshold cycle number (ΔC_T) in different tissues was tested using ANOVA. Post-hoc Tukey tests were also

21

performed. Changes in the expression were calculated using the expression $2^{-\Delta\Delta CT}$ (Schmittgen and Livak 2008).

5' RACE

RNA ligase-mediated rapid amplification of cDNA ends (RLM-RACE) was performed to identify the TSS for *Dntf-2r* in *D. simulans* using the First choice RLM-RACE kit from Ambion (Applied Biosystems/Ambion, Austin, TX). One µg of RNA extracted from adult males was used for 5' RACE in *D. simulans*; Figure 2). The gene-specific outer primer 5' CCGTTGGGCTTCAGCAAAAAGAT 3' and inner primer 5'CATCGCATTTTAGTCTTCCAAGGACG 3' were used along with the 5'RACE outer and inner primers provided by the manufacturer, respectively.

DNA samples and sequencing

To obtain divergence data for the regulatory region of *Dntf-2r*, DNA was extracted from a single fly using the Puregene DNA purification system from QIAGEN (QIAGEN, Valencia, CA). The region was amplified by PCR from genomic DNA using oligoprimers 5'TGCAGGGCGCATTGTTCAG3' and 5'CATACGCCTGCCAATACGAGT3' to amplify *Dntf-2r* in strains of *D. simulans*, *D. sechellia*, and *D. mauritiana*. PCR products were sequenced directly after purification (QIAGEN PCR purification kit; QIAGEN, Valencia, CA) on an ABI automated DNA sequencer using fluorescent DyeDeoxy terminator reagents (Applied Biosystems, Foster City, CA).

Preparation of the constructs, transformation and imaging

Dntf-2r, Pros28.1A *and flipped β2-UE1 constructs*

The complete *Dntf-2r* coding region and variable lengths of the upstream

regions were amplified from genomic DNA and cloned into the plasmid pEGFP1

(U55761; Clontech, Mountain View, CA) to put the *Dntf-2r* in frame with the EGFP

gene and generate a fluorescent fusion protein. The regions of these clones containing

different lengths of the putative regulatory region, 5'UTR of *Dntf-2r*, *Dntf-2r* - EGFP

fused coding regions and the SV40 polyadenylation site were further cloned into the

P element Drosophila transformation vector – pCaSpeR 4 (X81645) and used for fly

transformation.

The 14 bp β2-UE1-like sequence of *D. melanogaster* was replaced by the

orthologous 14 bp of *D. yakuba* to understand if this sequence drives testis-specific

expression. The region of *Dntf-2r* was amplified using the forward primer

5'CACC**GACCGGCTAGCGGC**CGACACAAACGAGAGGATATTTG3' that added

the *D. yakuba* sequence (bold) to the *D. melanogaster* region and the reverse primer

5´TTTAGTTCAAGTATATACGGGGTA3'. The PCR product was cloned with

topoisomerase-catalyzed reaction using pENTRY™ Directional TOPO cloning kit

(pENTR/D-TOPO Cloning Kit, Invitrogen catalog # K240020, Invitrogen, Carlsbad,

CA). The forward primer added the necessary recombination site to the PCR product

for directional cloning into the entry clone using the Gateway system. Colonies where

screened by colony PCR using primers flanking the gene followed by sequencing of

the positive clones. Miniprep for a good clone was performed, making this clone our master Entry clone for recombination into the UASt-GFP destination vector. To perform the recombination into the destination vector we used the Gateway LR Clonase Enzyme Mix (Invitrogen, Carlsbad, CA). Transformation was performed using Library Efficiency DH5alpha competent cells and the colonies where then sequenced by PCR to find positive clones. One good clone was chosen after sequencing and sent for transformation.

Another construct was made that included the described regulatory regions of the *β2-tubulin* gene and its 5'UTR (Michiels, et al. 1989) but in which the β2-UE1 element was flipped. This region was introduced upstream of EGFP in the pCaSpeR 4 transformation vector and used for fly transformation.

PCR primers were designed to amplify an increasingly shorter upstream region of *Pros28.1A* from a previous *Pros28.1A* –EGFP pCaSpeR 4 transformation vector (Sorourian and Betran 2010). The products containing different lengths of the *Pros28.1A* putative regulatory region, 5'UTR of *Pros28.1A* in front of EGFP and the SV40 polyadenylation site were cloned in the pCaSpeR 4 transformation vector and used for fly transformation.

These clones were sent for transformation into w[1118] flies to Genetic Services Inc. (Cambridge, MA). Several independent transformants were obtained for every

construct (See Results). More details about the primers designed and cloning procedures are given as Supplementary Materials and Methods.

Imaging

Tissues from transformant flies were dissected in Ringer solution and images were taken using an Olympus BX51TRF florescent microscope setting the UV exposure time manually at a level established by examining the control strain (w^{1118}; provided by Genetic Services Inc., Cambridge, MA). Controlling the time allowed us to fine tune for any autoflorescence in the original white mutant stock and enabled us to compare the level of green fluorescence across preparations and lines.

In situ hybridization

In situ hybridization of whole testis of young males of the Besançon strain was performed to detect the transcript of *Dntf-2r* following the protocol described by Morris et al. (Morris, et al. 2009) with some changes. We used single-stranded DNA probes instead of RNA probes. To denature the probe, it was heated at 100ºC for 10 minutes. The prehybridization and hybridization steps were carried out at 45ºC. The overnight incubation at 4ºC was done with 0.1% BSA in PBST and the anti-digoxygenin antibody. After the color development, the reactions were stopped with PBST washes (four times for 10 minutes each). After the four washes, all the PBST was removed and 30% glycerol in PBST was added for 30 minutes, followed by 50% glycerol, and, finally, 70% glycerol was added.

The probe labeling was performed using DIG-High Prime DNA Labeling and Detection Starter Kit I (Roche, Indianapolis, IN). A sense oligoprimer (5'CTTTTTTTCGGATCGGAAACTCAATCGTACCCCGTATATACTTGAACTA AAATGTCTCTGAATCTGCAGT3') allowed us to label the complementary DNA strand using the random priming procedure. This procedure was designed to detect the *Dntf-2r* transcript specifically as the oligo corresponds to a 5'UTR region that is different from the four transcripts of the parental gene, *Dntf-2-RA, Dntf-2-RB, Dntf-2-RC* and *Dntf-2-RE*.

Motif discovery

We downloaded the expression profile for the whole *D. melanogaster* gene set from Chintapalli and collaborators (Chintapalli, et al. 2007), available from FlyAtlas (http://www.flyatlas.org). We considered a gene to be transcribed testis-specifically when it was only up-regulated in testis and downregulated not in any other tissue (as in (Chintapalli, et al. 2007). A total of 1,245 genes were selected as testis-specific, including *Dntf-2r, Pros28.1A* and *Pros28.1B*. The region 300-bp upstream of the TSS of every gene in the *D. melanogaster* genome was downloaded from dmel_r.5.24. A sequence containing the β2UE1 of the *β2 tubulin* gene (AAATCGTAGTAGCCTATT) and a 13 bp testis-specific putative motif of *Pros28.1A* (see below) were BLASTed against these sequences. GLAM2 (Frith, et al. 2008) was used to build a Position Scoring Matrix (PSSM) from the BLAST hits, which was later given to the program GLAM2SCAN in order to scan the whole genome and find copies of the motifs. We used the top 26,000 motif matches found

by GLAM2SCAN in the genome to be sure we included as many potential motif matches as possible in our analyses. Those matches have a score value of 6.21 or higher. Matches located in upstream sequences of testis-specific genes (positive motifs) were used to build another PSSM that is used in the next iteration. This procedure allowed us to refine the PSSM and test for enrichment of the motif in testis-specific genes given an arbitrary score. The process continued for five iterations, i.e., until motif enrichment in testis-specific genes compared to non-testis-specific genes was no longer observed.

SUPPLEMENTARY MATERIAL

Supplementary Materials and Methods, Supplementary Results, a Supplementary Table and 8 Supplementary Figures and Supplementary Figure Legends are available for this publication.

LITERATURE CITED

Bai Y, Casola C, Betran E. 2008. Evolutionary origin of regulatory regions of retrogenes in Drosophila. BMC Genomics 9:241.

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.

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

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

Brosius J. (20035767 co-authors). 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115-134.

Caporilli S, Yu Y, Jiang J, White-Cooper H. 2013. The RNA export factor, NXT1, is required for tissue specific transcriptional regulation. PLOS Genet 9:e1003526.

Chen S, Ni X, Krinsky BH, Zhang YE, Vibranovski MD, White KP, Long M. 2012. Reshaping of global gene expression networks and sex-biased gene expression by integration of a young gene. EMBO J 31:2798-2809.

Chintapalli VR, Wang J, Dow JA. 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet 39:715-720.

Dai H, Yoshimatsu TF, Long M. 2006. Retrogene movement within- and between-chromosomes in the evolution of Drosophila genomes. Gene 385:96-102.

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

Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive Gene Traffic on the Mammalian X Chromosome. Science 303:537-540.

Fitzgerald PC, Sturgill D, Shyakhtenko A, Oliver B, Vinson C. 2006. Comparative genomics of Drosophila and human core promoters. Genome Biol 7:R53.

Frith MC, Saunders NF, Kobe B, Bailey TL. 2008. Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput Biol 4:e1000071.

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

Han MV, Hahn MW. 2012. Inferring the history of interchromosomal gene transposition in Drosophila using n-dimensional parsimony. Genetics 190:813-825.

Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. 2008. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLOS Genet 4:e1000106.

Hiller M, Chen X, Pringle MJ, Suchorolski M, Sancak Y, Viswanathan S, Bolival B, Lin TY, Marino S, Fuller MT. 2004. Testis-specific TAF homologs collaborate to control a tissue-specific transcription program. Development 131:5297-5308.

Kaessmann H. 2010. Origins, evolution and phenotypic impact of new genes. Genome Res 20:1313-1326.

Kalmykova AI, Nurminsky DI, Ryzhov DV, Shevelyov YY. 2005. Regulated chromatin domain comprising cluster of co-expressed genes in Drosophila melanogaster. Nucleic Acids Res 33:1435-1444.

Lankenau S, Corces VG, Lankenau DH. 1994. The Drosophila micropia retrotransposon encodes a testis-specific antisense RNA complementary to reverse transcriptase. Mol Cell Biol 14:1764-1775.

Lemeunier F, Ashburner MA. 1976. Relationships within the melanogaster species subgroup of the genus Drosophila (Sophophora). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. Proc R Soc Lond B Biol Sci 193:275-294.

Lemos B, Landry CR, Fontanillas P, Renn SCP, Kulathinal R, Brown KM, Hartl DL. 2008. Evolution of Genomic Expression. In: Pagel M, Pomiankowski A, editors. Evolutionary Genomics and Proteomics. Sunderland: Sinauer Associates. p. 81-118.

Li VC, Davis JC, Lenkov K, Bolival B, Fuller MT, Petrov DA. 2009. Molecular evolution of the testis TAFs of Drosophila. Mol Biol Evol 26:1103-1116.

Long M, Langley CH. (93219840 co-authors). 1993. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260:91-95.

Loppin B, Lepetit D, Dorus S, Couble P, Karr TL. 2005. Origin and neofunctionalization of a Drosophila paternal effect gene essential for zygote viability. Curr Biol 15:87-93.

Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564-567.

Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, Wilson RJ. 2013. FlyBase: improvements to the bibliography. Nucleic Acids Res 41:D751-757.

Meisel RP, Han MV, Hahn MW. 2009. A Complex Suite of Forces Drives Gene Traffic from Drosophila X Chromosomes. Genome Biol. Evol. 1:176-188.

Michiels F, Gasch A, Kaltschmidt B, Renkawitz-Pohl R. (89356626 co-authors). 1989. A 14 bp promoter element directs the testis specificity of the Drosophila beta 2 tubulin gene. EMBO J 8:1559-1565.

Mikhaylova LM, Boutanaev AM, Nurminsky DI. 2006. Transcriptional regulation by Modulo integrates meiosis and spermatid differentiation in male germ line. Proc Natl Acad Sci U S A 103:11975-11980.

Morris CA, Benson E, White-Cooper H. 2009. Determination of gene expression patterns using in situ hybridization to Drosophila testes. Nat Protoc 4:1807-1819.

Negre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, Feng X, Ahmad K, Russell S, White RA, et al. 2010. A comprehensive map of insulator elements for the Drosophila genome. PLOS Genet 6:e1000814.

Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL. (99075334 co-authors). 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature 396:572-575.

Ohler U, Liao GC, Niemann H, Rubin GM. 2002. Computational analysis of core promoters in the Drosophila genome. Genome Biol 3:RESEARCH0087.

Potrzebowski L, Vinckenbosch N, Marques A, Chalmel F, Jegou B, Kaessmann H. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol. 6:e80.

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

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

Quezada-Diaz JE, Muliyil T, Río J, Betran E. 2010. Drcd-1 related: a positively selected spermatogenesis retrogene in Drosophila. Genetica 138:925-937.

Sassone-Corsi P. 2002. Unique chromatin remodeling and transcriptional regulation in spermatogenesis. Science 296:2176-2178.

Schmittgen TD, Livak KJ. 2008. Analyzing real-time PCR data by the comparative C(T) method. Nat Protoc 3:1101-1108.

Schrider DR, Navarro FC, Galante PA, Parmigiani RB, Camargo AA, Hahn MW, de Souza SJ. 2013. Gene copy-number polymorphism caused by retrotransposition in humans. PLOS Genet 9:e1003242.

Schrider DR, Stevens K, Cardeno CM, Langley CH, Hahn MW. 2011. Genome-wide analysis of retrogene polymorphisms in Drosophila melanogaster. Genome Res 21:2087-2095.

Sorourian M, Betran E. 2010. Turnover and lineage-specific broadening of the transcription start site in a testis-specific retrogene. Fly (Austin) 4:3-11.

Vibranovski MD, Lopes HF, Karr TL, Long M. 2009. Stage-Specific Expression Profiling of Drosophila Spermatogenesis Suggests that Meiotic Sex Chromosome Inactivation Drives Genomic Relocation of Testis-Expressed Genes. PLoS Genetics 5:e1000731.

Vibranovski MD, Zhang Y, Long M. 2009. General gene movement off the X chromosome in the Drosophila genus. Genome Res 19:897-903.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A 103:3220-3225.

Wang W, Zhang J, Alvarez C, Llopart A, Long M. (20416306 co-authors). 2000. The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in Drosophila melanogaster. Mol Biol Evol 17:1294-1301.

White-Cooper H. 2010. Molecular mechanisms of gene regulation during Drosophila spermatogenesis. Reproduction 139:11-21.

Yang J, Porter L, Rawls J. 1995. Expression of the dihydroorotate dehydrogenase gene, dhod, during spermatogenesis in Drosophila melanogaster. Mol Gen Genet 246:334-341.

Yang J, Ramos E, Corces VG. 2012. The BEAF-32 insulator coordinates genome organization and function during the evolution of Drosophila species. Genome Res 22:2199-2207.

Yuan X, Miller M, Belote JM. (97032933 co-authors). 1996. Duplicated proteasome subunit genes in Drosophila melanogaster encoding testes-specific isoforms. Genetics 144:147-157.

Zhang Z, Hambuch TM, Parsch J. 2004. Molecular Evolution of Sex-Biased Genes in Drosophila. Mol Biol Evol 21:2130-2139.

FIGURE LEGENDS

Figure 1. A. *Dntf-2r*-EGFP fusion constructs used for transformation in *D. melanogaster*. The drawn line depicts the regions included in every construct. It is drawn proportional to the region and its length varies from -151 to -4 bp upstream of the TSS. The white box depicts 5'UTR and coding region (CDS) of *Dntf-2r* (not drawn to scale). +1 denotes the start of transcription (i.e., TSS) of the gene fusion. Green box depicts EGFP coding region (not drawn to scale). Fluorescence results due to the testis-specific expression of the fusion protein are summarized. See Supplementary Fig. 4 for fluorescent pictures of all independent insertions. All constructs except the last two show strong fluorescence only in testis revealing that a testis-specific element is located between -78 and -51 bp from the TSS. B. In situ results revealing the transcription in testis of *Dntf-2r* are shown and coincide with the observed expression of the fusion protein. C. The construct including *D. yakuba* β2-UE1-like region (Figure 2) is outlined and the fluorescence results summarized. See Supplementary Fig. 5 for fluorescent pictures of all independent insertions. D. The construct including β2-UE1 flipped and the *β2-tubulin* region (Figure 2) is outlined and the fluorescence results summarized. See Supplementary Fig. 8 for fluorescent pictures of all independent insertions.


Figure 2. A. Percent identity of the upstream region of several testis-specific genes to the known 14-bp β2UE1 motif and the 7-bp motif with quantitative effect of the *β2 tubulin* gene. Identical bases are shown in bold. Negative numbers show the location

of both elements as distance to the transcription start site. For example β2UE1 motif of the *β2 tubulin* gene in *D. melanogaster* spans the base pairs from -51 to -38 and the 7-bp motif with quantitative effects in the same gene and species spans the base pairs from -32 to -26. This information is consolidated from our study and work of others (Michiels, et al. 1989; Lankenau, et al. 1994; Yang, et al. 1995; Nurminsky, et al. 1998). Yellow highlighted region shows the sequence required for testis-specific expression of *Dntf-2r* in *D. melanogaster*. B. *Dntf-2r* upstream region in the species where the gene is present and the same region in species where the gene is absent. Similar sequences were observed in the syntenic region in distantly related species suggesting that a testis-specific regulatory region existed before *Dntf-2r* insertion. The nucleotide state at the ancestral node was inferred when the same nucleotide was observed in species where *Dntf-2r* is present and in species where *Dntf-2r* is absent and not inferred (?) otherwise. We highlight in green the region used from *D. yakuba* to make a construct to show the current regulatory potential of the 14 bp in this species.

Figure 3. A. Constructs carrying the *Pros28.1A* upstream region and 5'UTR in front of EGFP used for transformation in *D. melanogaster* and a summary of their level of fluorescence are shown. Line depicts the region included in the construct and its length varies from 189 to 0 bp (not drawn to scale). White box depicts 5'UTR of *Pros28.1A* (not drawn to scale). +1 denotes the start of transcription of the gene. Green box depicts EGFP coding region (not drawn to scale). Zigzag line denotes the 23 bp region that was mutagenized. Lack of EGFP expression in the shortest

construct and the mutagenic construct of a 23-bp region is observed revealing that the regulatory regions needed for testis expression includes the regions between -23 and 0 bp. B. Alignment of the upstream region of *Pros28.1A* of *D. melanogaster* with the same region in other species that contain the gene. Yellow highlighted region (49 bp) shows the sequence that was included in the 23 bp construct in front of EGFP and it also depicts how much of the 5'UTR was included in all the constructs (Figure 3 A). The 23-bp region required for testis-specific expression is shown in bold and the *D. melanogaster* transcript is shown in italics.

Figure 1

A.

	Independent insertions	Expression in testis
-151	8	++
-101	7	+
-77	8	+
-50	7	−
-4	5	−
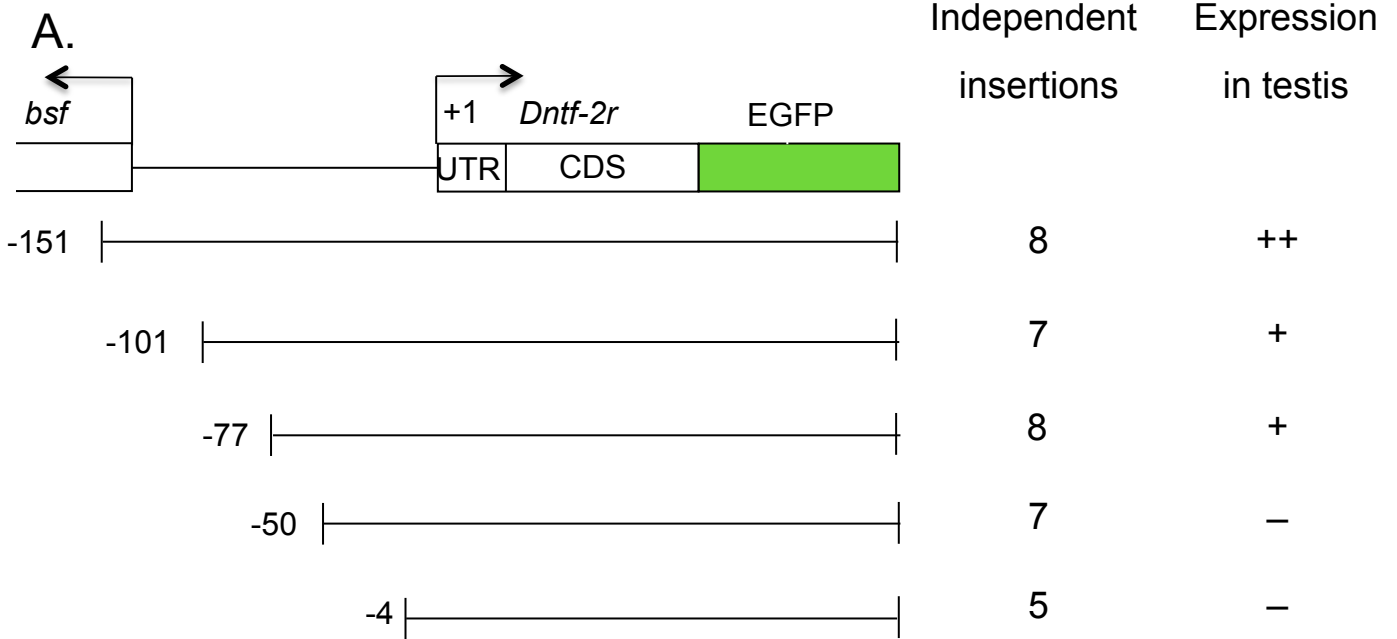
B.
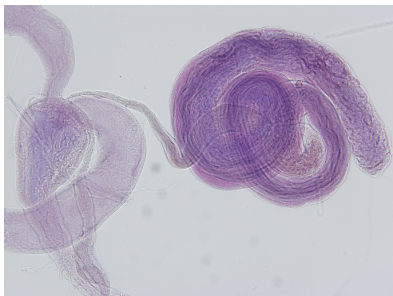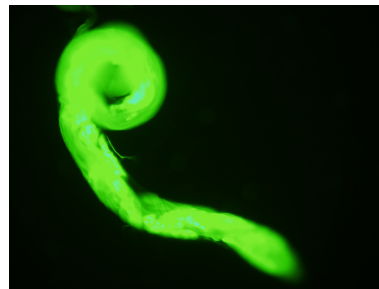
Dntf-2r in situ   Fusion protein expression (-151 bp construct)

C.
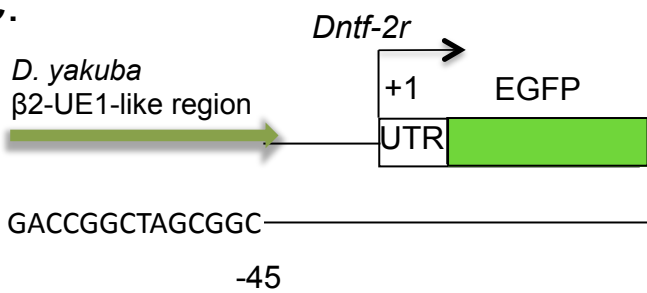
D. yakuba β2-UE1-like region

GACCGGCTAGCGGC
-45

Independent insertions	Expression in testis
18	+

D.

D. melanogaster β2-UE1 flipped

TAGGCTACTACGAT
-37
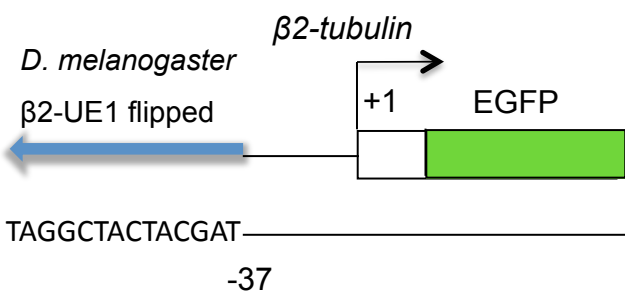
Independent insertions	Expression in testis
10	+

Figure 2

A.

	β2-UE1 (14 bp)	Quantitative element (7 bp)

β2 tubulin
```
D. melanogaster   ATC-G--TAGTA-GCCTATT--------TGTGAACATT  (-51|-38/-32|-26)
D. hydei          ATC-G--CAGTA-GTCTAAC--------TAGGGATATT  (-51|-38/-32|-26)
```

Sdic
```
D. melanogaster   ATC-G--TAGTGTGCCTTTG--------GGGGAA-ATT  (-68|-54/-48|-43)
                     (86%)                    (86%)
```

Dntf-2r
```
D. mauritiana     ATCAG--CT-TA-GCGGTGACACAAACGAGAGGATATT
D. simulans       ATCAG--CT-TA-GCGGTGACACAAACGAGAGGATATT  (-58|-45/-31|-25)
D. sechellia      ATCAG--CT-TA-GCGGTGACACAAACGACAGGATATT
D. melanogaster   GTATCAG--CT-TA-GCGGTGACACAAACGAGAGGATATT  (-75|-62/-48|-42)
                     (64%)              (100%)
```

Dihydroorotate dehydrogenase
```
D. melanogaster   GCC-G--TAGTA-A-CTATATATCGCATGAGTTTTCCT  (-50|-38)
                     (71%)
```

microcopia
```
D. hydei          TTC-GTCTAGTA-GGTCAT----------TTGAATTCA  (-24|-9/-5|+2)
                     (71%)                    (57%)
```

B.

Dntf-2r present
```
D. mauritiana    GTATCAGCT-TA-GCGGTGACACAAACGAGAGGATATT
D. simulans      GTATCAGCT-TA-GCGGTGACACAAACGAGAGGATATT
D. sechellia     GTATCAGCT-TA-GCGGTGACACAAACGACAGGATATT
D. melanogaster  GTATCAGCT-TA-GCGGTGACACAAACGAGAGGATATT
```

Dntf-2r absent
```
D. yakuba        GTGACCGGC-TA-GCGGCGACACAAGCCAAGGGATATT
D. teissieri     GTGACCCGGCTA-GCGGCGACACAAGCCAAGGGATATT
D. erecta        AATCTAGCC-CA-GCGGATACACAAACCGAGGGATATT

D. eugracilis    CTTTTTGTC-AG-TAATCGATTTTACAGGCCCAGTGTG

Ancestral inferred  GT??CAGC?-TA-GCGG?GACACAAAC?A??GGATATT
```

Retroduplication
Ancestral

Figure 3

A. *D. melanogaster* constructs

Pros28.1A +1 EGFP

UTR

Construct	Independent insertions	EGFP expression in testis
-189	7	++
-80	6	++
-46	9	++
-23	10	+
-46	5	-
0	6	-

B.

23 bp region

```
Pros28.1A
D. mauritiana   GAAAAATTCATATTGTTTCAACAA-------------AATAACTGCGTTGCAAGCAAAC-ATTATATTCTG-TTGCCTGCTAACTAATCCAAAGGATG
D. sechellia    GAAAAATTCATATTGTTTCAACAA-------------AATAACTGC-TGGCAAGCAAAC-ATTATATTCTGGTTGCCTGCTAACTAATCCAAAGGATG
D. simulans     GAAAAATTCATATTGTTTCAACAA-------------AATAACTGC-TTGCAAGCAAAC-ATTATATTTTG-TTGCCTGCTAACTAATCCAAAGGATG
D. melanogaster GAAAAATTCATATTGTTTCAAGTA-------------AATAACT-TGTCACAAGCAAATTGTTGTTTTGTT-TTGCCTGCTAACTAACCCAAAGGATG
D. yakuba       GAAAAACTAATTACTTTTCAACTTTACTTTTCAAATAACTAACTGG-TCGCAAGCAAGT-ATAGATTGCT-CTGCCCACTAGCAAGTCCCAAGGATG
D. santomea     GAAAAACTAATTACTTTTCAACTTTACTTTTCAAATAACTAACTGG-TCGCAAGCGAGT--ATAGATTGCT-CTGCCCACTAGCAAGTCCCAAGGATG
D. teissieri    GAAAAACTCATAACTTTTCAACCA-------------AACTAACTGG-TCGCAAGCGAGT--ATAGATTGCT-CTGCCCACTAGCAAGCCCCAAGGATG
D. erecta       GAAAACCTAATTTTTCAACCA-AG-------------AATATTCAATATTCAAGCGAG--AGTATATTGTT-CTGCCCACTAACTAATCCCAAGGATG
                *****  *  **         *  *              *  **      *****        *    **     ****  *** *  *  **  *******
```