6654–6661 *Nucleic Acids Research, 2005, Vol. 33, No. 20*
doi:10.1093/nar/gki969

A retrocopy of a gene can functionally displace the source gene in evolution

Aleksey N. Krasnov[1,2], Maria M. Kurshakova[1], Vasily E. Ramensky[3], Pavel V. Mardanov[1], Elena N. Nabirochkina[1,2] and Sofia G. Georgieva[1,2,3,*]

[1]Russian Academy of Sciences, Institute of Gene Biology, 119334 Moscow, Russia, [2]Centre for Medical Studies, University of Oslo, 119334 Moscow, Russia and [3]Russian Academy of Sciences, Engelhardt Institute of Molecular Biology, 119991, Vavilova 32, Moscow, Russia

Received August 3, 2005; Revised October 10, 2005; Accepted November 2, 2005

ABSTRACT

The *e(y)2* gene of *Drosophila melanogaster* encodes the ubiquitous evolutionarily conserved co-activator of RNA polymerase II that is involved in transcription regulation of a high number of genes. The *Drosophila e(y)2b* gene, paralogue of the *e(y)2* has been found. The analysis of structure of the *e(y)2*, *e(y)2b* and its orthologues from other species reveals that the *e(y)2* gene derived as a result of retroposition of the *e(y)2b* during *Drosophila* evolution. The mRNA-derived retrogenes lack introns or regulatory regions; most of them become pseudogenes whereas some acquire tissue-specific functions. Here we describe the different situation: the *e(y)2* retrogene performs the general function and is ubiquitously expressed, while the source gene is functional only in a small group of male germ cells. This must have resulted from retroposition into a transcriptionally favorable region of the genome.

INTRODUCTION

Enhancers of yellow [*e(y)*] is a group of genetically and functionally related genes for proteins involved in transcriptional regulation. They have been originally isolated in *Drosophila melanogaster* in genetic screen aimed to find mutations influencing the activator-dependent transcription (1). It is important that the weak mutations of the *e(y)* genes that do not influence the viability of flies proved to be lethal in compound, suggesting that these genes have overlapping and/or redundant functions (1). In our previous studies, we have shown that *e(y)1* encodes TAF9, a subunit of both TFIID and the TFTC complexes (2) while the *e(y)3* gene encodes a multidomain co-activator of RNA polymerase II (Pol II). The E(y)3 protein has a general role in regulation of transcription in both euchromatin and heterochromatin (3).

In our previous studies we have also demonstrated that the *e(y)2* encodes a novel ubiquitous transcription factor of Pol II, 101 amino acids in length, and is highly conserved in evolution (4). It has been shown that the weak $e(y)2^{u1}$ mutation influences the phenotype of weak mutations in the *yellow*, *white*, *cut* and *scute* genes (1).The weak $e(y)2^{u1}$ mutation also causes the multiple disturbances in the fly phenotype (4) while the strong mutation is lethal. According to this fact we conclude that E(y)2 is involved in transcription regulation of wide number of genes during fly development. Importantly the cDNA of human homologue was found in different tissues. It demonstrates that both *Drosophila* and human proteins are ubiquitously expressed (4).

Drosophila E(y)2 co-activates transcription on chromatin template and is a component of a large multiprotein complex that contains TAF9 (4). Recently Sus1, yeast counterpart of *Drosophila* E(y)2, has been found and shown to be the component of SAGA histone acetyltransferase complex (5). It suggests that E(y)2 is also a component of GCN5 HAT-containing complex that was previously detected in *D.melanogaster* (6–8). Sus1 has been also shown to play an important role in transcription coupled mRNA export (5). Overall obtained data suggest that E(y)2/Sus1 protein is an essential player of transcription machinery of eukaryotes.

Comparatively large portion of eukaryotic genomes is represented by retrogenes, created as a result of reverse transcription of mRNA. These genes lose introns and original regulatory elements that usually are not co-transferred. Therefore the vast majority of retrogenes become silent and owing to the absence of selection pressure accumulate mutations, gaining the features of pseudogenes. However in few cases retrogenes may stay transcriptionally active. It may happen in two cases: either the retrocopy is inserted under the preexisting promoter sequence, or integration of retrocopy creates the

*To whom correspondence should be addressed. Tel: +7 095 135 9731; Fax: +7 095 135 1405; Email: sonjag@molbiol.edu.ru

sequences of the novel promoter (9,10). Some retrogenes were shown to encode functional proteins (11,12). Their expression is usually tissue-specific, in particular testis-specific, while the majority of source genes are ubiquitously expressed (9,13,14).

In our paper we describe the *D.melanogaster* paralogue of *e(y)2* hereafter referred as the *e(y)2b*. While the *e(y)2b* has three exons and shares the exon–intron structure with the *e(y)2* homologues from other species [hereafter referred as the *e(y)2/sus1*], the *e(y)2* contains only one exon and is flanked by direct repeats. Obtained data suggest that the *e(y)2* gene originated as the result of retroposition of the source *e(y)2b* copy. As the *e(y)2/sus1* genes from different species are ubiquitously expressed, one can suggest the existence of the same expression pattern for the *e(y)2b*. However the expression of the *e(y)2b* gene is testis-specific while the *e(y)2* is ubiquitously expressed. The *e(y)2b* encodes the protein of 95 amino acids which is unable to replace E(y)2 functionally. Thus the retrogene becomes actively transcribed and takes over the functions of the source gene, while the latter is converted into the gene expressed only in male germ cells.

MATERIALS AND METHODS

Cloning of the *e(y)2b*, finding of orthologues and promoter prediction

The *e(y)2b* gene (CG14612) was found in GeneBank (NM_144053). To confirm the predicted structure of the *e(y)2b*, we performed RT–PCR on mRNA isolated from an adult fly. The product was cloned and sequenced. ClustalW 1.83 alignment (http://www.genebee.msu.su/clustal) of the cDNA sequence of the *e(y)2b* obtained by RT–PCR to the genomic sequence proved that the *e(y)2b* exon–intron structure coincides with the predicted one. Sequences of the *e(y)2* homologues have the following accession numbers in GeneBank: *D.melanogaster e(y)2* gene (AF173294); *D.melanogaster e(y)2b* gene (AE003672); mRNA of the *e(y)2* gene from *D.melanogaster* (AF173295), *Mus musculus* (AF173297), *Homo sapiens* (AF173296) and *Saccharomyces cerevisiae* (AY278445); mRNA of the *e(y)2b* gene from *D.melanogaster* (NM_144053). The *e(y)2* and the *e(y)2b* genes from *Drosophila pseudoobscura* were found in the Human Genome Sequencing Center database (http://www.hgsc.bcm.tmc.edu/projects/drosophila): *e(y)2*—contig1045_contig3832, *e(y)2b*—contig268_contig6582.

Transgenic constructs

For rescue experiments, the cDNAs of the *e(y)2* and the *e(y)2b* were cloned in CaSpeR-3 vector under the *Su(Hw)* promoter (15) (*P{w*$^+$, *Su(Hw)_e(y)2}* and *P{w*$^+$, *Su(Hw)_e(y)2b}* constructs, respectively).To assess the regulatory regions of the *e(y)2* and *e(y)2b*, *LacZ* was cloned under genomic sequences of *e(y)2* (−408 to +48) (+1 start of coding region) or the *e(y)2b* (−950 to +48) in CaSpeR-AUG-betagal vector.

Antibodies and protein extracts

Antibodies against the peptide MTINKETGTDPDPGYKPC specific for E(y)2b (Ab1) and antibodies against His-tagged E(y)2b (Ab2) were raised in rabbits. Both of them were affinity-purified. Ab2 were further depleted against recombinant E(y)2 that was coupled to CNBr-activated Sepharose to avoid cross-reaction on western blots. The polyclonal antibodies against E(y)2 described previously (4) were conversely depleted against recombinant E(y)2b before western blotting. Testes were dissected in phosphate-buffered saline (PBS), followed by immediate adding of SDS sample buffer. Preparation of embryonic extracts was described previously (4). SDS–PAGE (15%) was used to identify E(y)2 on western blots.

mRNA preparation and northern blot analysis

Isolation of total RNA from *Drosophila* embryos, larvae, pupae or imagoes and northern hybridization was performed as described (16). The poly(A)$^+$ RNA (1.5 mg) was loaded per lane of agarose gel. Membranes were exposed to a Storage Phosphor Screen and developed on a Cyclone Storage Phosphor System (Packard Instrument Company). Total cellular RNA from mouse tissues was extracted with Trizol (Invitrogen) following the manufacture's recommendation. The probes corresponding to complete cDNA of the *e(y)2*, *e(y)2b* and *e(y)2/sus1* of mouse were used for northern hybridization.

β-Galactosidase activity assay

To study *LacZ* expression, testes were obtained from 5-day-old transgenic males, fixed in 2% glutaraldehyde in PBS for 30 min, and stained overnight at 37°C in X-gal-containing buffer as described elsewhere (17).

Genetic crosses and *P*-element-mediated transformation

Cultivation of flies, genetic crosses and isolation of the $e(y)2^{u1}$ mutation were described previously (1). The $e(y)2^{u1}$ mutation was maintained in $y^2w^1e(y)2^{u1}/FM4$ strain. The level of y^2 expression was measured as described previously (1). The number of inserted copies was determined by Southern blot analysis using the *P*-element sequence as a probe. The constructs were injected into y^1w^1 preblastoderm embryos as described elsewhere (18,19). The analysis of the rescuing capacities of transgenes was performed in the $e(y)2^{u1}$ y^2w^7 males bearing either *P{w*$^+$, *Su(Hw)_e(y)2b}* construct, or the *P{w*$^+$, *Su(Hw)_e(y)2}* control construct.

RESULTS

Search for the *e(y)2b* gene and its exon–intron structure analysis

The sequence of the E(y)2b protein of *D.melanogaster* was found by Blast search in the RefSeq protein database (20). The sequence identity between E(y)2 and its paralogue E(y)2b (also known as CG14612-PA) is 41% over 77 aligned residues. The comparison of the genomic sequence and cDNA demonstrated that the *e(y)2b* contains two introns (Figure 1A). The exon–intron structure of the gene was also confirmed by cloning and sequencing. As was shown previously, the *e(y)2* is a one-exon gene (4).

With the help of Blast search in species genomes at FlyBase (21), the single-exon and three-exon homologues of the *e(y)2* were also found in completely sequenced genome of *D.pseudoobscura* [GA13559-PA and GA13111-PA, with 74

Figure 1. The comparison of the *e(y)2*/*sus1* genes and proteins from different species. (**A**) The structure of the *e(y)2* and the *e(y)2b* genes of *D.melanogaster* and *D.pseudoobscura*. Exons are indicated as dark boxes. Numbers show lengths of corresponding exon and intron in nucleotides. (**B**) The sequences of upstream (upper lines) and downstream (lower lines) direct repeats flanking the *e(y)2* of *D.melanogaster* and *D.pseudoobscura*. The nucleotides identical to consensus are highlighted. (**C**) Multiple alignment of e(y)2, e(y)2b and their homologues from other species. Grey rectangles denote the intron shadows. The three proteins at the bottom have no introns.

and 46% sequence homology to e(y)2, respectively]. The comparison of the *e(y)2b* of *D.melanogaster* with its orthologue in *D.pseudoobscura* revealed their similar exon–intron structure (Figure 1A).

The *e(y)2* gene of *D.melanogaster* is flanked by 17 bp direct repeats beginning at positions −12 and +528 nt relative to the transcription start site. The repeats are the hallmarks of retroposition (9) and were also found in *D.pseudoobscura* (Figure 1B). This result together with the revealed one-exon structure of the *e(y)2* gene suggests that the *e(y)2* is a retroposed copy of the *e(y)2b*. As it could be expected for the source gene and its retrocopy, the *e(y)2b* and the *e(y)2* are located on different chromosomes (84B6 and 10C7) and the sequences surrounding these genes are completely different too.

The direct repeats also look more conserved than the surrounding sequences. We have found the traces of these repeats in several other *Drosophilae* species but they look more diverged there. This conservation may be explained by insertion of retroposed copy in the functional sequence that still remained functionally significant after the duplication. This could probably happen as the gene density in the region of insertion is high. The recent studies in the comparative genomics field show that the high degree of between-species conservation of non-coding sequences with yet unknown function is much more frequent than has been expected before, with conserved regions occupying up to 53% of the *D.melanogaster* genome (22). The sequences of direct repeats contain several CA dinucleotides which resemble the microsatellites. This does not contradict the suggestion that insertion lead to the duplication of this particular sequence.

The exon–intron structure of the *e(y)2b* is similar to homologous genes in other species

The Blast search of *D.melanogaster* E(y)2 amino acid sequence against protein RefSeq and non-redundant GenBank CDS databases reveals 19 protein hits with *E*-value $<10^{-4}$ (Supplementary Table 1). The human, dog and mouse proteins have identical amino acid sequences. The pairwise sequence identities of the query and its homologues from *Drosophila rerio*, *H.sapiens*, *Arabidopsis thaliana* and *S.cerevisiae* are 50, 48, 43 and 33%, respectively. For *e(y)2b*, the corresponding figures are 40, 38 and 36%, with no detectable similarity to Sus1 protein from *S.cerevisiae*. The higher degree of E(y)2 conservation agrees with the hypothesis that it has functionally replaced the source protein and the latter is therefore subject to lower selection pressure.

Of 19 found hits, we have selected 9 proteins from eight species that have no indication 'predicted', 'unknown', 'hypothetical' or 'conceptual translation' in the Entrez Protein records. Figure 1C shows the multiple sequence alignment of these confirmed proteins produced by ClustalW (23).

With the help of BLAT (24) and tblastn (20) tools, all obtained protein sequences have been aligned to chromosomes

from corresponding organisms. This analysis shows the number and positions of introns in the corresponding genes. The number of introns varies from 0 to 4; the intron shadows are marked with grey rectangles above the protein sequences on Figure 1C. It is interesting that besides the *e(y)2* genes from two *Drosophila* species, the *Ciano intestinalis* gene is also intronless, as seen from the BLAT alignment with the chromosome sequence. However, the study of this gene is beyond the scope of this paper. One can see that except for few cases the intron positions are quite conservative suggesting that the intron-containing genes are othologous to each other and exist in many species.

Search for the orthologous protein sequences in the insect genomes

With the help of VISTA genome alignment tools (25), sequences of E(y)2 and E(y)2b have been found in eight available genomes of *Drosophilae* species (Supplementary Table 2). E(y)2b sequences from *Drosophila mojavensis*, *Drosophila virilis* and *Drosophila ananassae* could not be reliably identified owing to the insufficient quality of genome alignments. The phylogenetic tree built for the fly sequences and orthologues described in the previous paragraph is shown in the Supplementary Figure. One can see that the observed E(y)2 and E(y)2b sequence relationship reflects the taxonomic similarity generally assumed for *Drosophilae* species (26).

In order to check whether the products of the *e(y)2b* retroposition event can be observed in available genomes from the insect lineage, we have performed the Tblastn search of *D.melanogaster e(y)2*, *e(y)2b* and *Arabidopsis gambiae* orthologue of the *e(y)2b* (ENSANGP00000025782) sequences against *A.gambiae* genomic seguences. Besides the trivial recovery of the two-exon orthologue of the *e(y)2b*, no hits were obtained at the *E*-value cutoff relaxed up to 0.1. The table of syntenic regions between *A.gambiae* and *D.melanogaster* (27) suggests no partners neither for the *e(y)2* nor for its neighbouring genes.

The genomic sequences of *Arabidopsis melifera* have been studied in the similar manner with Tblastn and VISTA browser that currently contains the *D.melanogaster–A.melifera* whole genome alignment. The analysis revealed only the sequence that can be recognized as the second exon of the *e(y)2b* orthologue, with no indication of existence of retroposed copy. The first exon is missing from the Tblastn hit table, owing to its short length and possible divergence from the query sequences.

The absence of the retroposed copy in the mosquito and honey bee genomes suggests that the retroposition probably occurred after the insect speciation but before the emergence of various *Drosophilae* species.

The *e(y)2b* expression is tissue specific

The *e(y)2* gene is actively expressed in almost all cells of *Drosophila* (4). The same was found for the *e(y)2*/*sus1* in human (4). Northern analysis revealed the presence of mRNA corresponding to *e(y)2*/*sus1* in different mouse tissues (Figure 2A). We further investigated expression pattern of the *e(y)2b* and compared it with that of the *e(y)2*. While mRNA of the *e(y)2* was detected at all stages of development at approximately the same level, the expression of the *e(y)2b* was found only in males and pupae (Figure 2B). The expression in pupae was low, increasing in late pupae and becoming high in adult males. The study of the *e(y)2b* expression in different tissues revealed the presence of its mRNA only in testes (Figure 2C). Thus, the increase of the level of the *e(y)2b* transcription is

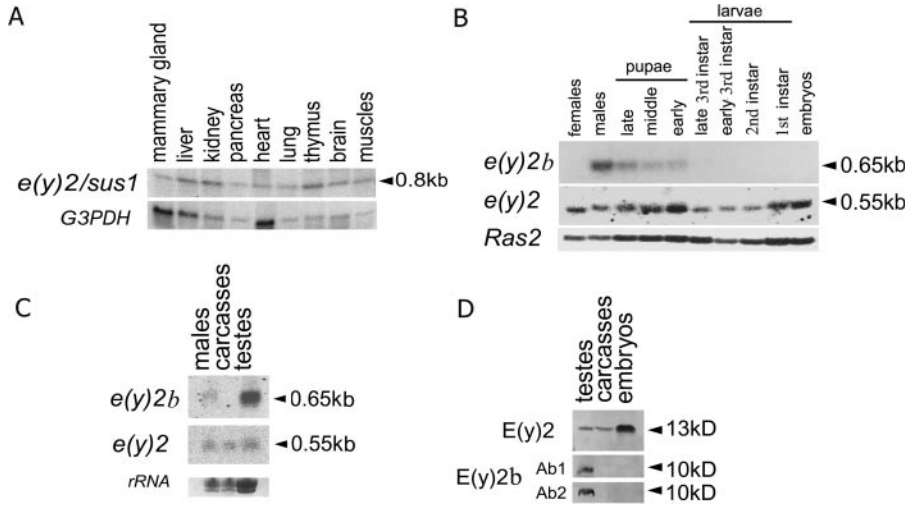

Figure 2. The expression of the *e(y)2b* is tissue-specific while the *e(y)2* is ubiquitously expressed. (**A**) Northern hybridization of RNA (15 μg per lane) isolated from different mouse tissues with probe for the mouse *e(y)2*/*sus1*. *G3PDH* was used as internal gel loading control. (**B**) Northern hybridization of poly(A)$^+$ RNA (3 μg per lane) isolated at different stages of development of *D.melanogaster* with probes for the *e(y)2b* and the *e(y)2*. *ras2* was used as internal gel loading control. (**C**) Northern blot hybridization of RNA from adult males, males without germ line cells (carcasses) and testes with *e(y)2b* and *e(y)2* probes. The rRNA (stained with etidium bromide) was used as gel loading control (lower panel). (**D**) Western blot hybridization of protein extracts from testes, carcasses and embryos with antibodies specific either for E(y)2 or for E(y)2b. Different anti-E(y)2b antibodies raised either against short peptide specific for E(y)2b (Ab1) or against the recombinant protein (Ab2) were used. The antibodies against the recombinant E(y)2b gave the cross-reaction with E(y)2. Thus aiming to use them for development of western blot, we depleted them against recombinant E(y)2. Vice-versa the antibodies against E(y)2 were depleted against recombinant E(y)2b.

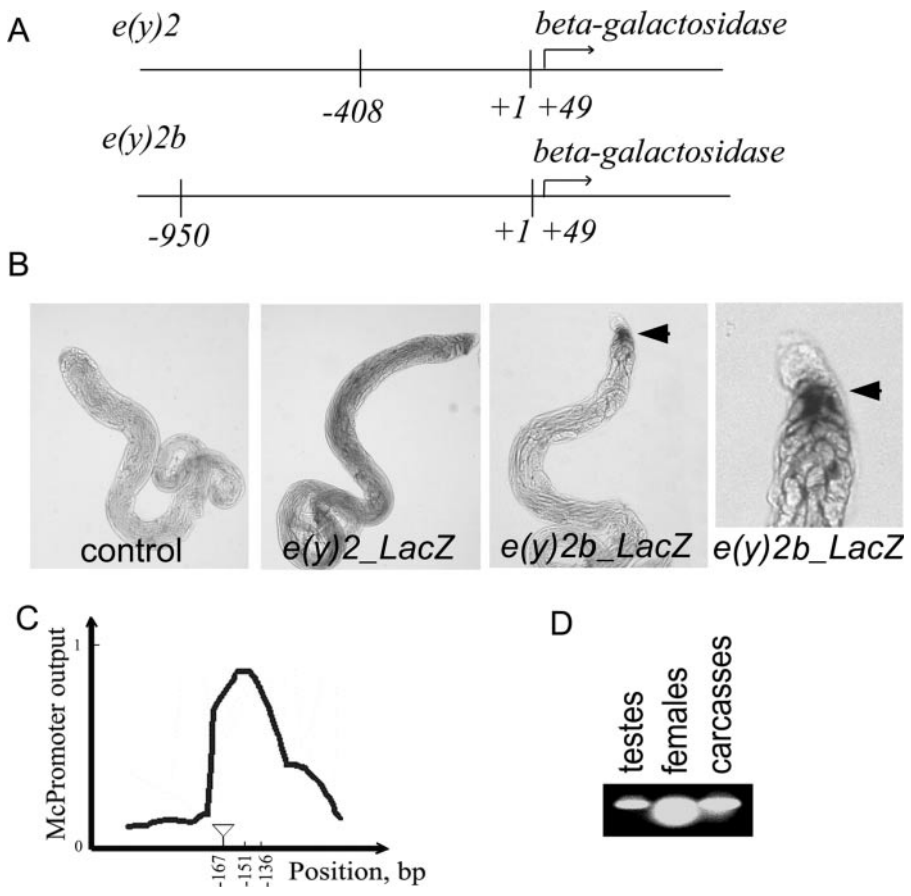

Figure 3. The expression of the *e(y)2b* is restricted to primary spermatocytes. (**A**) Schematic representation of transgenes carrying *LacZ* under the *e(y)2* (*e(y)2_LacZ*) or the *e(y)2b* (*e(y)2b_LacZ*) regulatory sequences. (**B**) X-gal staining of testes dissected from transgenic and from control wild-type males. Right panel represents the magnified version of the previous one. Arrowhead indicates the area of *LacZ* expression in the *e(y)2b_LacZ* flies. (**C**) Prediction of the *e(y)2* promoter *in silico* by McPromoter program (15). The highest probability is determined to 15 nt at position from −37 to −23 nt, relative to transcription start (from −151 to −136 relative to the beginning of the open reading frame). Triangle indicates the site of *Stalker* insertion. (**D**) The ubiquitous transcription of the *e(y)2b* driven by *Su(Hw)* promoter in transgenic flies shown by RT–PCR.

observed as testes develop from early pupae stage to adult male. The *e(y)2* mRNA was also found in testes, but at the same time it was present at similar level in other tissues (Figure 2C).

To investigate, whether the *e(y)2b* is a protein-coding gene two different polyclonal antibodies were raised in rabbits: the first one was against short peptide specific for E(y)2b and the second one against the whole recombinant protein. In preparations from *Drosophila* testes, both antibodies detected the band of molecular weight corresponding to that calculated from the amino acid sequence. The protein was absent in preparations from males, lacking germ line and embryonic extract. This fact confirmed our expectations that the detected protein is E(y)2b. However, E(y)2 was detected in all samples (Figure 2D).

The upstream regions of the *e(y)2* and the *e(y)2b* genes provide their expression pattern: while the *e(y)2* is ubiquitously expressed, the expression of the *e(y)2b* is restricted by primary spermatocyte stage

Two transgenes that contained β-galactosidase (*LacZ*) gene under the control of sequences containing either the *e(y)2* or the *e(y)2b* regulatory region were constructed (Figure 3A). They were used to compare the expression patterns of the *e(y)2* and the *e(y)2b*.

The β-galactosidase gene driven by the *e(y)2* promoter was ubiquitously expressed both in testes (Figure 3B) and in other tissues of transgenic flies (data not shown). In contrast, *LacZ* under the *e(y)2b* regulatory region displayed an expression pattern restricted to a narrow zone in the apical part of testis which rapidly dwindled thereafter (Figure 3B). The observed staining coincides with the beginning of primary spermatocyte stage of differentiation of male germ cells. The β-galactosidase expression was neither detected in the tip of testis in mitotically dividing germ cells nor in gonial cells. Thus the *e(y)2b* gene retained its expression only in a small group of differentiating male germ cells.

The obtained data demonstrate that the upstream region of the *e(y)2* provides its ubiquitous expression. The McPromoter program (28) designed to search for *Drosophila* core promoters predicted the promoter around position −30 nt, upstream of the *e(y)2* transcription start (Figure 3C). The essential role of this region is confirmed by the fact that the insertion of *Stalker* mobile element (−53) profoundly

Table 1. The results of experiments on rescue of the *e(y)2*^{u1} mutation by transgenes expressing E(y)2 or E(y)2b proteins

Genotype	Number of strains studied	Level of pigmentation[a]		Viability[b]%	Distorted tergites[c]%
		Head bristles	Thorax bristles		
e(y)2^{u1}		3	2	72	14
e(y)2^{u1}; *P{w*$^+$, *Su(Hw)_e(y)2}*	3	5	5	93–98	0
e(y)2^{u1}; *P{w*$^+$, *Su(Hw)_e(y)2b}*	5	3	2	69–74	12–16
e(y)2^{u1}; *P{w*$^+$, *Su(Hw)_e(y)2b}* × 2	1	3	2	74	15

The phenotype of *e(y)2*^{u1} *y*^{2}*w*^{1} males of different strains bearing the transgene was analysed.
[a]Evaluated in 3 to 5-day-old males developing at 25°C, ranked on a scale from 0 (pigmentation of *y*^{1} flies) to 5 (pigmentation of *y*$^+$ flies).
[b]Percentage of surviving transgenic males versus FM4 males. At least 200 males were scored for each transgenic strain.
[c]Percentage of transgenic males displaying the mutated phenotype versus normal transgenic males. At least 200 transgenic males were scored for each strain.

decreased the level of the *e(y)2* transcription in the *e(y)2*^{ul} mutated strain of flies (4).

The E(y)2b protein is unable to replace E(y)2 functionally

Finally, we checked whether the E(y)2b protein could perform the functions of E(y)2. The *e(y)2*^{u1} mutation that was mentioned has several morphological manifestations. In particular, it causes the decrease of *yellow* gene expression in head and thorax bristles of flies of *y*^{2} allele. It also causes female sterility, decreases viability and has weak but diverse effect on fly morphology: short stocky body, separated wings and eyes with irregular facets (1). The *e(y)2*^{u1} males demonstrate abnormal development of anal plates. We tested if ubiquitous expression of the *e(y)2b* is able to rescue the mutated *e(y)2*^{u1} allele.

The cDNA of the *e(y)2b* was cloned under ubiquitously expressing *Su(Hw)* promoter in *CaSpeR*-3 vector (*P{w*$^+$, *Su(Hw)_e(y)2b}*) and as the result five lines bearing the transgene in different sites of genome were obtained. The *P{w*$^+$, *Su(Hw)_e(y)2}* construct containing the *e(y)2* cDNA under the *Su(Hw)* promoter in *CaSpeR*-3 was used as the control.

Testing different tissues of the transgenic *e(y)2*^{u1}; *P{w*$^+$, *Su(Hw)_e(y)2b}* flies by RT–PCR for the presence of the *e(y)2b* mRNA confirmed that transgene was ubiquitously expressed (Figure 3D). However careful analysis demonstrated that neither of independent insertions of *P{w*$^+$, *Su(Hw)_e(y)2b}* was able to complement any features of mutated phenotype of the *e(y)2*^{u1} strain (Table 1). Even the introduction of two copies of the construct in mutated flies did not have any effect. On the contrary, four obtained insertions of the *P{w*$^+$, *Su(Hw)_e(y)2}* control transgene completely rescued the wild-type phenotype of the *e(y)2*^{u1} strain. Thus, the divergence between E(y)2b and E(y)2 is strong enough to disable E(y)2b to replace E(y)2 functionally.

DISCUSSION

Here we describe the *e(y)2b* gene of *D.melanogaster*, a paralogue of the earlier discovered *e(y)2* gene. Obtained data demonstrate that in contrast to the ubiquitously expressed *e(y)2*, the *e(y)2b* displayed expression pattern restricted to primary spermatocytes. The beginning of the *e(y)2b* expression coincides with the onset of primary spermatocyte stage and rapidly dwindles thereafter.

Primary spermatocyte stage of differentiation of male germ cells is characterized by a high level of gene expression. Most of expressing genes determine the following meiotic divisions and spermatid differentiation program (29). Several testis-specific homologues of general transcription factors of *Drosophila* and vertebrates were found to express at spermatocytes suggesting that transcription program specific for male germ cells may utilize alternative Pol II transcription machinery (14,30). Recently five testis-specific homologues of genes encoding *Drosophila* TAFs were found to express in primary spermatocytes (30). They collaborate in the control of transcription of subset of target genes involved in spermatid differentiation. Our previous data demonstrated E(y)2 to be present in a multiprotein complex and to interact with several components of TFIID both genetically and in co-immunoprecipitation experiments (4). It is suggested that the tissue-specific homologues of different components of transcription machinery cooperate to drive the specific transcription program at primary spermatocytes.

The obtained data imply that the *Drosophila e(y)2* gene is mRNA-derived retrocopy of the intronized *e(y)2b* that has lost both introns as a result of retroposition. Indeed it is considered that the majority of intronless genes present in eukaryotic genomes have been generated by retroposition [for review see (9)]. As the *e(y)2* homologue in *Anopheles* has only one intron (Supplementary Table 1) one may speculate that the common ancestor of the mosquito and the fly possessed the one-intron gene that we observe in mosquito genome. In this case, the observed *Drosophila* paralogues should have been emerged as a result of gene duplication with subsequent loss of intron in one copy and intron acquisition in the other one. This complicated scenario is doubtful and contradicts the intron position conservation across *e(y)2b* orthologues. The presence of direct repeats flanking the *e(y)2* sequence is another independent argument supporting the hypothesis of retroposition.

The ubiquitous expression of the *e(y)2/sus1* in vertebrates suggests the *Drosophila e(y)2b* had once been also ubiquitously expressed but then lost its general function in the process of evolution. One may speculate on how the source gene turned into male specific and its expression became restricted to primary spermatocytes. The probable explanation is that initially *e(y)2b* had two promoters, one responsible for ubiquitous expression and the other one utilized in primary spermatocytes. This is a common case for many ubiquitously expressed genes (28). The ubiquitous promoter could further deteriorate because of accumulation of mutations in this region.

Usually retrogenes lacking regulatory sequences are rapidly converting into pseudogenes. Most of the known mRNA-derived retrogenes are tissue-specific, while the source gene is ubiquitously expressed (12). The recently described

example is the functional replacement of human ubiquitously expressed general transcription factor TAF1 by its retroposed homologue TAF1L expressed in testes (14). Here we describe the opposite case when descendant copy became more essential than the ancestor gene.

The comparison of amino acid sequences of different E(y)2 homologues demonstrates that E(y)2 protein is more evolutionary conserved than E(y)2b. Indeed our experiments demonstrate that the source gene has 'evolved' so profoundly that its protein product has become incapable of the original function. These data are in agreement with the observation that *Drosophila* genes with male-biased expression had significantly faster rates of evolution than genes with female-biased or unbiased expression (31). This fact, together with the lower selection pressure acting on the initial gene after the 'successful' retroposition, may explain the degree of divergence of two copies comparable with that observed in comparison with distant eukaryotic species. The observed absence of the retroposed copy of the *e(y)2b* in the *A.gambiae* and *A.melifera* genomes indicates that the retroposition may have occurred after the insect speciation but before the emergence of various *Drosophilae* species.

We suggest that the processed mRNA-derived *e(y)2* copy after retroposition has been inserted in transcriptionally active region probably under the control of a strong resident promoter or other positive regulatory elements. It could sustain the *e(y)2* transcription at a higher level than that of the source gene. Therefore, the *e(y)2* retrogene came under strong selection pressure, while *e(y)2b* escaped it and genetically drifted much faster. As a result, the *e(y)2* has functionally displaced the source gene. This is an example whereby a retrogene takes over the functions of the source gene in evolution. Interestingly, the predicted *Drosophila* gene for ubiquitously expressed TAF5 (FBgn0010356) also has a single exon whereas its testis-specific homologue contains four introns (13), so such supersedence may be not a rare case.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Pavel Georgiev and Mikhail Gelfand for helpful discussion and Tatiana Loutchnick for help in preparation of the manuscript. This work was supported by two Cellular and Molecular Biology grants from the Russian Academy of Sciences, INTAS #211, and CRDF #RB1-2349-MO-02. Funding to pay the Open Access publication charges for this article was provided by Cellular and Molecular Biology grant from RAS.

Conflict of interest statement. None declared.

REFERENCES

1. Georgiev,P.G. (1994) Identification of mutations in three genes that interact with zeste in the control of white gene expression in *Drosophila melanogaster*. *Genetics*, **138**, 733–739.
2. Soldatov,A., Nabirochkina,E., Georgieva,S., Belenkaja,T. and Georgiev,P. (1999) TAFII40 protein is encoded by the *e(y)1* gene: biological consequences of mutations. *Mol. Cell. Biol.*, **19**, 3769–3778.
3. Shidlovskii,Y.V., Krasnov,A.N., Nikolenko,J.V., Lebedeva,L.A., Kopantseva,M., Ermolaeva,M.A., Ilyin,Y.V., Nabirochkina,E.N., Georgiev,P.G. and Georgieva,S.G. (2005) A novel multidomain transcription coactivator SAYP can also repress transcription in heterochromatin. *EMBO J.*, **24**, 97–107.
4. Georgieva,S., Nabirochkina,E., Dilworth,F.J., Eickhoff,H., Becker,P., Tora,L., Georgiev,P. and Soldatov,A. (2001) The novel transcription factor e(y)2 interacts with TAF(II)40 and potentiates transcription activation on chromatin templates. *Mol. Cell. Biol.*, **21**, 5223–5231.
5. Rodriguez-Navarro,S., Fischer,T., Luo,M.J., Antunez,O., Brettschneider,S., Lechner,J., Perez-Ortin,J.E., Reed,R. and Hurt,E. (2004) Sus1, a functional component of the SAGA histone acetylase complex and the nuclear pore-associated mRNA export machinery. *Cell*, **116**, 75–86.
6. Georgieva,S., Kirshner,D.B., Jagla,T., Nabirochkina,E., Hanke,S., Schenkel,H., de Lorenzo,C., Sinha,P., Jagla,K., Mechler,B. *et al.* (2000) Two novel *Drosophila* TAF_{II}s have homology with human $TAF_{II}30$ and are differentially regulated during development. *Mol. Cell. Biol.*, **20**, 1639–1648.
7. Muratoglu,S., Georgieva,S., Papai,G., Scheer,E., Enunlu,I., Komonyi,O., Cserpan,I., Lebedeva,L., Nabirochkina,E., Udvardy,A. *et al.* (2003) Two different *Drosophila* ADA2 homologues are present in distinct GCN5 histone acetyltransferase-containing complexes. *Mol. Cell. Biol.*, **23**, 306–321.
8. Kusch,T., Guelman,S., Abmayr,S.M. and Workman,J.L. (2003) Two *Drosophila* Ada2 homologues function in different multiprotein complexes. *Mol. Cell. Biol.*, **23**, 3305–3319.
9. Brosius,J. (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene*, **238**, 115–134.
10. Mighell,A.J., Smith,N.R., Robinson,P.A. and Markham,A.F. (2000) Vertebrate pseudogenes. *FEBS Lett.*, **468**, 109–114.
11. Persson,K., Heby,O. and Berger,F.G. (1999) The functional intronless S-adenosylmethionine decarboxylase gene of the mouse (Amd-2) is linked to the ornithine decarboxylase gene (Odc) on chromosome 12 and is present in distantly related species of the genus Mus. *Mamm. Genome*, **10**, 784–788.
12. Lingenfelter,P.A., Delbridge,M.L., Thomas,S., Hoekstra,H.E., Mitchell,M.J., Graves,J.A. and Disteche,C.M. (2001) Expression and conservation of processed copies of the RBMX gene. *Mamm. Genome*, **12**, 538–545.
13. Hiller,M.A., Lin,T.Y., Wood,C. and Fuller,M.T. (2001) Developmental regulation of transcription by a tissue-specific TAF homolog. *Genes Dev.*, **15**, 1021–1030.
14. Wang,P.J. and Page,D.C. (2002) Functional substitution for TAF(II)250 by a retroposed homolog that is expressed in human spermatogenesis. *Mol. Genet.*, **11**, 2341–2346.
15. Kim,J., Shen,B., Rosen,C. and Dorsett,D. (1996) The DNA-binding and enhancer-blocking domains of the *Drosophila* suppressor of Hairy-wing protein. *Mol. Cell. Biol.*, **16**, 3381–3392.
16. Maes,M. and Messens,E. (1992) Phenol as grinding material in RNA preparations. *Nucleic Acids Res.*, **20**, 4374.
17. Gvozdev,V.A., Aravin,A.A., Abramov,Y.A., Klenov,M.S., Kogan,G.L., Lavrov,S.A., Naumova,N.M., Olenkina,O.M., Tulin,A.V. and Vagin,V.V. (2003) Stellate repeats: targets of silencing and modules causing *cis*-inactivation and trans-activation. *Genetica*, **117**, 239–245.
18. Spradling,A.C. and Rubin,G.M. (1982) Transposition of cloned P elements into *Drosophila* germ line chromosomes. *Science*, **218**, 341–347.
19. Rubin,G.M. and Spradling,A.C. (1982) Genetic transformation of *Drosophila* with transposable element vectors. *Science*, **218**, 348–353.
20. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
21. Drysdale,R.A. and Crosby,M.A., and The FlyBase Consortium (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33**, D390–D395.
22. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

23. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
24. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
25. Couronne,O., Poliakov,A., Bray,N., Ishkhanov,T., Ryaboy,D., Rubin,E., Pachter,L. and Dubchak,I. (2003) Strategies and tools for whole-genome alignments. *Genome Res.*, **13**, 73–80.
26. Bergman,C.M., Pfeiffer,B.D., Rincon-Limas,D.E., Hoskins,R.A., Gnirke,A., Mungall,C.J., Wang,A.M., Kronmiller,B., Pacleb,J., Park,S. *et al.* (2002) Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.*, **3**, RESEARCH0086.
27. Zdobnov,E.M., von Mering,C., Letunic,I., Torrents,D., Suyama,M., Copley,R.R., Christophides,G.K., Thomasova,D., Holt,R.A., Subramanian,G.M. *et al.* (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science, **298**, 149–159.
28. Ohler,U., Liao,G.C., Niemann,H. and Rubin,G.M. (2002) Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.*, **3**, RESEARCH0087.
29. Fuller,M.T. (1998) Genetic control of cell proliferation and differentiation in *Drosophila* spermatogenesis. *Semin. Cell Dev. Biol.*, **9**, 433–444.
30. Hiller,M., Chen,X., Pringle,M.J., Suchorolski,M., Sancak,Y., Viswanathan,S., Bolival,B., Lin,T.Y., Marino,S. and Fuller,M.T. (2004) Testis-specific TAF homologs collaborate to control a tissue-specific transcription program. *Development*, **131**, 5297–5308.
31. Zhang,Z., Hambuch,T.M. and Parsch,J. (2004) Molecular evolution of sex-biased genes in *Drosophila*. *Moll. Biol. Evol.*, **21**, 2130–2139.