Insulator proteins contribute to expression of gene loci repositioned into heterochromatin in the course of *Drosophila* evolution

Sergei Yu. Funikov¹, Alexander P. Rezvykh¹,², Dina A. Kulikova³, Elena S. Zelentsova¹, Lyubov N. Chuvakova¹, Venera I. Tyukmaeva⁴, Irina R. Arkhipova⁵ and Michael B. Evgen’ev¹*¹

¹Engelhardt Institute of Molecular Biology of Russian Academy of Sciences, Moscow, Russia
²Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia
³Koltzov Institute of Developmental Biology of Russian Academy of Sciences, Moscow, Russia
⁴Department of Biological and Environmental Science, University of Jyväskylä, 40014, Finland
⁵Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts, USA

*Corresponding author – Michael B. Evgen’ev, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russia

E-mail: misha672011@yahoo.com

Key words: *Drosophila*, heterochromatin, gene expression, molecular evolution, insulators, BEAF-32

Abstract

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27
Pericentric heterochromatin in *Drosophila* is generally composed of repetitive DNA forming a transcriptionally repressive environment. Nevertheless, dozens of genes were embedded into pericentric genome regions during evolution of *Drosophilidae* lineage and retained functional activity. However, factors that contribute to “immunity” of these gene loci to transcriptional silencing remain unknown. Here, we investigated molecular evolution of the essential *Myb* and *Ranbp16* genes. These protein-coding genes reside in euchromatic loci of chromosome X in *D. melanogaster* and related species, while in other studied *Drosophila* species, including evolutionary distant ones, they are located in genomic regions highly enriched with the remnants of transposable elements (TEs), suggesting their heterochromatic nature and location. The promoter region of *Myb* exhibits a conserved structure throughout the *Drosophila* phylogeny and carries motifs for binding of chromatin remodeling factors, including insulator BEAF-32, regardless of eu- or heterochromatic surroundings. Importantly, BEAF-32 occupies not only the promoter region of *Myb* but is also found in the vicinity of transcriptional start sites (TSS) of *Ranbp16* gene as well as in a wide range of genes located in the contrasting chromatin types in *D. melanogaster* and *D. virilis*, denoting the boundary of the nucleosome-free region available for RNA polymerase II recruitment and the surrounding heterochromatin. We also find that along with BEAF-32, insulators dCTCF and GAF are enriched at the TSS of heterochromatic genes in *D. melanogaster*. Thus, we propose that insulator proteins contribute to gene expression in the heterochromatic environment and, hence, facilitate the evolutionary repositioning of gene loci into heterochromatin.

**Author summary**

Heterochromatin in *Drosophila* is generally associated with transcriptional silencing. Nevertheless, hundreds of essential genes have been identified in the pericentric heterochromatin
of *Drosophila melanogaster*. Interestingly, genes embedded in pericentric heterochromatin of *D. melanogaster* may occupy different genomic loci, euchromatic or heterochromatic, due to repositioning in the course of evolution of *Drosophila* species. By surveying factors that contribute to the normal functioning of the relocated genes in distant *Drosophila* species, i.e. *D. melanogaster* and *D. virilis*, we identify certain insulator proteins (e.g. BEAF-32) that facilitate the expression of heterochromatic genes in spite of the repressive environment.

### Introduction

Eukaryotic genomes are packaged in chromatin consisting of DNA and associated proteins. Typically, chromatin can be divided into two basic forms, euchromatin and heterochromatin (1). These types of chromatin are distinguished by several distinctive properties, including DNA sequence composition, specific histone modifications and binding proteins, nuclear and chromosomal localization, and frequency of meiotic recombination (1, 2). One of the major subtypes of heterochromatin in *Drosophila* is marked by heterochromatin protein 1 (HP1a) and di- or trimethylated H3K9 (3, 4). This subtype of heterochromatin covers large genomic segments primarily around centromeres and, in association with the protein POF (painting of fourth), the entire dot chromosome 4 in *D. melanogaster* (3-5). Pericentric heterochromatin is mainly composed of repetitive sequences, including remnants of various transposable elements (TEs), satellite DNAs and other repeats (6). A distinctive feature of heterochromatin is the ability to silence euchromatic genes placed within heterochromatic environment due to chromosomal inversions or transposition events, a phenomenon called position effect variegation (PEV) (7-12). Transcriptional silencing of euchromatic genes in PEV is mediated by spreading of heterochromatin-associated marks HP1a and H3K9me3 across the gene loci transferred to heterochromatin (8, 10).
Despite the repressive environment, dozens of essential genes were identified in the pericentric heterochromatin of *D. melanogaster* (13-16). Interestingly, genes embedded in pericentric heterochromatin in *D. melanogaster* may occupy distinct genomic loci, euchromatic and heterochromatic, in other *Drosophila* species (17). For instance, two adjacent genes *RpL15* and *Dbp80* located in the pericentric region of chromosome 3L in *D. melanogaster* reside in a euchromatic region in *D. pseudoobscura* (18). A similar pattern was demonstrated for genes *light* and *Yeti* located in pericentric regions in *D. melanogaster*, while in *D. virilis* they are found within euchromatin on the same chromosomal elements (19, 20). Recently, it was shown that most of the pericentric genes found at both arms of chromosome 2 of *D. melanogaster* are located in euchromatic loci in the *D. virilis* genome (21).

However, although repositioning of genes between euchromatin and heterochromatin during genome evolution is not unusual in the *Drosophilidae* lineage, the “immunity” of heterochromatic genes to the transcriptionally repressive environment remains paradoxical and unexplained. Thus, it is not clear whether these loci have undergone adaptation to heterochromatic environment or had some intrinsic properties permitting local adaptation. Previously, it was shown that molecular organization of promoter regions is largely conserved between heterochromatic and euchromatic genes, indicating that adaptation to heterochromatin probably does not require major changes in regulatory sequences (20). However, expression of heterochromatic genes requires the methylated H3K9 mark (22, 23), and the ability to repositioning during evolution is a characteristic feature of gene clusters that show close association with HP1a protein (21).

Chromatin insulator elements and associated proteins were originally defined by their ability to protect transgenes from PEV, exerting a block to *cis* spreading of a chromatin state (24, 25). To date, a set of insulators have been identified in *Drosophila*, including BEAF-32 (Boundary element associated factor of 32 kDa), dCTCF (*Drosophila* homolog of CTCF), Su(Hw) (Suppressor of hairy wing), Zw5 (Zeste-white-5), GAF (GAGA factor) and recently...
described Pita and ZIPIC (zinc-finger protein interacting with CP190) (26, 27). Numerous studies demonstrated that insulators are responsible for a vast number of genomic functions, including stimulation of gene transcription, enhancer-blocking and barrier insulation partitioning of eukaryotic genomes into independently regulated domains (27-29). Hence, one may hypothesize that gene loci capable of adaptation to heterochromatin probably share specific sites of insulation that ensure their expression in the repressive environment.

To address this issue, we investigated molecular evolution of Myb and Ranbp16 genes in the Drosophilidae lineage. Myb is an essential gene encoding a transcription factor involved in transition from G2 to M phase of the cell cycle (30, 31). Ranbp16 encodes a RanGTP-binding protein belonging to the importin-β superfamily and mediates translocation of proteins into the nucleus. Both genes are located in euchromatic loci of the D. melanogaster X-chromosome, while in other studied Drosophila species belonging to Sophophora and Drosophila subgenus they are found in genomic regions with a high density of repetitive DNA elements upstream, downstream and within introns, suggesting their location in pericentric heterochromatin. Regardless of the euchromatic or heterochromatic surroundings, the promoter region of Myb displays high sequence homology and stable structural organization among Drosophila species studied so far. We find that the conserved motifs in the promoter sequence of Myb serve as a binding site for the chromatin insulator protein BEAF-32 and transcriptional factor DREF (The DNA replication-related element (DRE)/DNA replication-related element-binding factor). Furthermore, we demonstrate that BEAF-32 occupies not only the promoter region of Myb in two evolutionarily distant species, D. melanogaster and D. virilis, but also, in cooperation with other insulators dCTCF and GAF, is present in the promoters of most studied heterochromatin genes described in D. melanogaster. Importantly, the evolutionary gene repositioning between euchromatin and the pericentric heterochromatin occurred with preservation of sites of insulation, keeping the binding of BEAF-32 in close proximity to the transcription start sites of these genes. Overall, we propose that insulator proteins, in particular BEAF-32, contribute to
expression of heterochromatic genes and make them predisposed for evolutionary repositioning into transcriptionally repressive genomic environments.

Results

Evolutionary repositioning of Myb and Ranbp16 genes between euchromatin and heterochromatin in Drosophila phylogeny

In order to determine whether Myb and Ranbp16 gene locations have been rearranged on the evolutionary timescale, we first mapped these genes onto genomic scaffolds of Drosophila species separated by evolutionary distances from 5 to 40 million years (32-36). These include species of the melanogaster group (D. melanogaster and D. yakuba) and the obscura group (D. persimilis and D. miranda), with both groups belonging to the Sophophora subgenus, along with the virilis group (D. virilis and D. novamexicana) and the repleta group (D. mojavensis and D. hydei) that belong to the Drosophila subgenus. Next, we performed comparative analysis of Myb and Ranbp16 genes, as well as the intergenic regions between these genes. As indicated in Fig. 1, the coding sequences of Myb (red nodes) and Ranbp16 (blue nodes) genes are highly homologous between the species studied (Fig. 1). However, the regions between Myb and Ranbp16 genes differ significantly among Drosophila species studied, exhibiting sequence homology mostly within the related groups. For instance, while Myb and Ranbp16 genes of D. melanogaster and D. yakuba are embedded within a large gene cluster, their orthologues in other Drosophila species reside in the genomic regions mostly occupied by repetitive sequences, including the introns of Ranbp16 gene (Fig. 1). Thus, one may conclude that Myb and Ranbp16 genes in the species belonging to the melanogaster group are located in euchromatin, in the region containing a large gene cluster. On the other hand, the localization of the studied genes in the environment packed with repetitive elements and apparently representing a pericentric heterochromatic region is typical for most other Drosophila species studied here including the
According to the integrated physical and cytogenetic map of Schaeffer et al. (37), as well as overall homology of the studied regions, *Myb* and *Ranbp16* genes are located at the chromosome X in all *Drosophila* species studied.

Next, we studied in more detail the genomic loci containing *Myb* and *Ranbp16* genes, focusing on two evolutionarily distant species, *D. melanogaster* and *D. virilis*, separated by 40 million years of evolution (32). The single copies of *Myb* and *Ranbp16* genes map to the chromosome X of *D. melanogaster* at the cytogenetic loci 13F14 and 14A1 of salivary gland polytene chromosomes, respectively (Fig. 2A). These regions are significantly less enriched with heterochromatic marks H3K9me3 and HP1a than telomeric and pericentric regions of the chromosome and, hence, represent typical euchromatin. As mentioned earlier, *Myb* and *Ranbp16* genes in *D. melanogaster* are located at a distance ~80 Kb from each other within a large protein-coding gene cluster, which includes only a few repetitive sequences (Fig. 2A).

To confirm that *Myb* and *Ranbp16* reside in a heterochromatic region in *D. virilis*, we performed DNA *in situ* hybridization on polytene chromosomes using the unique sequence of *Myb* gene as a probe. *In situ* hybridization indicates that *Myb* is located at the base of chromosome X near the chromocenter in *D. virilis* (Fig. 2B). Mapping analysis of *Myb* and *Ranbp16* genes on the genomic scaffolds of *D. virilis* reveals that both genes reside in one scaffold_13050 at a distance ~120 Kb from each other (Fig. 2C). In contrast to *D. melanogaster*, the region between these genes in *D. virilis* does not contain any other protein-coding genes and consists of remnants of TEs and other repeats diverged to varying degrees (Fig. S1). We used annotation of known genomic scaffolds of *D. virilis* made by Schaeffer et al. (37) to assign the proximal scaffold of *D. virilis* genome r1.06 to the centromeric region of chromosome X. According to specific marker genes, we retrieved scaffold_12970 and extended it with scaffold_13050 containing *Myb* and *Ranbp16* genes, keeping some space unassembled between these scaffolds (Fig. 2C). Enrichment profile of H3K9me3 clearly indicates that the putative euchromatin-heterochromatin border lies in the proximal 1Mb of scaffold_12970 (Fig. 2C).
whole scaffold_13050, including the region where *Myb* and *Ranbp16* are localized, is heavily occupied by the H3K9me3 mark in comparison with the most contiguous fragment of scaffold_12970 (Fig. 2C).

Taken together, our data indicate that the essential *Myb* and *Ranbp16* genes reside in contrasting chromatin regions, euchromatic or heterochromatic, in the species of the *Drosophila* genus studied so far.

Both *Myb* and *Ranbp16* genes are under purifying selection regardless of their heterochromatic or euchromatic localization

Next, we examined whether coding sequences of *Myb* and *Ranbp16* underwent negative (purifying) or positive selection during evolution and what was the impact of heterochromatic location on molecular evolution of these genes. To this end, we first estimated the number of base substitutions per site in their coding sequences (Fig. S2A, B), and then calculated the ratio of non-synonymous to synonymous substitutions (dN/dS) (Fig. S2C, D) using eight aforementioned representatives of the Sophophora and Drosophila subgenera, as well as the sequences from *Anopheles gambiae* as an outgroup. The results suggest that both *Myb* and *Ranbp16* genes are under purifying (negative) selection (dN/dS < 0.2 in all pairs of comparison) when present either in heterochromatin or euchromatin (Fig. S2).

We also performed multiple alignment of protein sequences of *Myb* and *Ranbp16* from Diptera species and their orthologs in mouse and human. The results show a high degree of conservation of both proteins even between dipteran and mammalian species (Fig. S3). The consensus sequences of aligned proteins are properly recognized as Helix-Turn-Helix (HTH) DNA-binding domain and Importin-β domain that characterize the described structure of *Myb* and *Ranbp16* proteins, respectively (Fig. S3).

Evolutionary changes in gene organization of *Myb* and *Ranbp16* in *Drosophila* species
To study how the structure of Myb and Ranbp16 genes has been changed in the course of evolution, resulting in their placement in different chromatin contexts, we compared the structure of these genes in eight Drosophila species.

As a starting point, we retrieved the available sequences of Myb and Ranbp16 genes from D. melanogaster, D. yakuba, D. persimilis, D. obscura, D. virilis, D. mojavensis and D. hydei from Flybase and NCBI databases. To date, full annotation of genes including 5’ and 3’UTR (untranslated region) of most Drosophila species, except for D. melanogaster, is absent. To resolve this issue and obtain the most complete sequence of studied genes, we used sequences of closely related species (for example, D. obscura sequences were used to resolve the D. persimilis gene structure, etc.). Only those alignments that mapped on the same DNA strand adjacent to the existing annotated sequence with E-value > e-80 were considered as relevant and used for reconstruction of gene structure. Also, 5’RACE (Rapid amplification of cDNA ends) was performed to determine the missing 5’UTR of Ranbp16 gene of D. virilis. Due to the lack of annotation, the sequences of Myb and Ranbp16 genes of D. novamexicana were reconstructed using the sequences of related D. virilis species and blastn algorithm. Gaps in the continuous alignment were considered as introns. All information, including accession numbers and genomic coordinates of the studied genes in Drosophila species, is listed in supplementary table S1.

Despite prolonged evolution, the structure of Myb gene was preserved virtually unchanged in terms of overall gene length and exon-intron organization (Fig. 3A). A single intron located in the protein-coding region of this gene is conserved both in position and approximate size in the species of both subgenera (Fig. 3A). However, sequence homology of this intron is conserved only within species belonging to one subgenus, Sophophora or Drosophila. This suggests that the intronic sequence per se probably does not affect gene regulation. Therefore, the heterochromatic Myb gene tends to be stable during evolution of the
Drosophila genus and retains the single intron of approximately the same size in different species.

In contrast to *Myb*, the structure of *Ranbp16* is more complex and demonstrates greater flexibility in terms of extension of total gene length by increasing the size of its introns (Fig. 3B). The length of *Ranbp16* gene in the species of melanogaster group (*D. melanogaster* and *D. yakuba*) is ~ 6700 bp and includes up to 15 introns not exceeding 407 bp in length. In the course of evolution, the gene structure of *Ranbp16* in other studied *Drosophila* species was expanded to varying degrees by multiple insertions of TEs. For example, the length of *Ranbp16* gene is approximately 100 Kb in *D. obscura*, 55 Kb in *D. virilis* and 40 Kb in *D. mojavensis* (Fig. 3B).

Interestingly, the introns prone to TE insertions increasing the length of the gene are strictly defined. They include the first, third and fifth introns from the TSS (transcriptional start site) of the gene, as well as a few introns at the distal end of the gene (Fig. 3B). The extension of introns located in the middle of the gene was not observed in any of the studied species, suggesting that these introns may have a structural or regulatory role in *Ranbp16* function.

Thus, the structure of heterochromatic genes may exhibit great flexibility in terms of overall gene length extension, if this extension does not disrupt gene activity.

**Promoter of Myb gene contains motifs for binding of chromatin remodeling proteins in Drosophila**

To reveal specific properties of gene loci allowing the essential *Myb* and *Ranbp16* genes to be actively transcribed in both euchromatin and heterochromatin, we studied the promoter region of *Myb* by searching for common and different motifs in *Drosophila* species. For this purpose, we expanded the list of analyzed species to 19 representatives of the Sophophora and 11 representatives of the Drosophila subgenera. As mentioned above, due to the absence of gene annotation for a range of studied species (*D. miranda*, *D. guanche*, *D. subobscura* and all virilis group species with the exception of *D. virilis*) the putative TSS was set as the first mapped
nucleotide of 5′UTR of *Myb* gene of related species. Due to the insertion of DAIBAM MITE element at the position 92 bp upstream of the TSS of *Myb* gene in *D. virilis*, the analyzed region was shortened to 100 nt for all species. Using the MEME suite (38) we were able to identify three motifs that are present in the promoter of *Myb* gene in all analyzed species of *Drosophila* (Fig. 4A). Search through OnTheFly (39) and REDfly v5.6 (40, 41) databases of known transcription factors and their binding sites indicates that the two highest-scoring motifs contain a potential binding site for insulator protein BEAF-32 and transcriptional factor DREF (Fig. 4B). The third motif shows limited homology to the binding site of ecdysone receptor co-activator Taiman and another undescribed factor (Fig. 4B).

**Insulator protein BEAF-32 is enriched in promoters of *Myb* and *Ranbp16* genes in both *D. melanogaster* and *D. virilis* species**

To demonstrate that the insulator protein BEAF-32 binds to the promoter region of *Myb* gene, we used ChIP-seq data to profile BEAF-32 occupancy in the gene loci containing *Myb* and *Ranbp16* genes in *D. melanogaster* and *D. virilis*. Furthermore, we applied ChIP-seq to profile RNA polymerase II (Pol II) distribution in the promoters of *Myb* and *Ranbp16* genes. Additionally, we used ATAC-seq data to correlate BEAF-32 and Pol II enrichment with the nucleosome-free conformation of chromatin in the promoters of these genes indicative of intensive transcription. Mapping of RNA-seq and H3K9me3 ChIP-seq data also provides valuable information regarding expression levels of these genes and the heterochromatic profile of the analyzed gene loci, respectively.

As seen in Fig. 5A, in *D. melanogaster* the locus containing *Myb* gene and adjacent *AlkB* gene is highly enriched with BEAF-32, with the peak summit of mapped BEAF-32 reads in the promoter of the *Myb* gene (Fig. 5A). The enrichment of Pol II and mapped ATAC-seq reads reside within the region starting from the peak of BEAF-32 binding to the TSS of *Myb*, indicating that BEAF-32 binds to DNA at the boundary of the *Myb* gene locus (Fig. 5A).
Likewise, the promoter of Ranbp16 gene of D. melanogaster is enriched with BEAF-32, located slightly upstream from the Pol II binding site and open chromatin region (Fig. 5B).

A similar pattern of BEAF-32 occupancy in the promoters of heterochromatic Myb and Ranbp16 genes is observed in D. virilis (Fig. 5C). However, the enrichment profile of BEAF-32 upstream of the TSS of Ranbp16 gene in D. virilis and D. melanogaster has one difference which is worth mentioning. In contrast to D. melanogaster, where BEAF-32 is enriched in close proximity to the TSS of Ranbp16, the binding of BEAF-32 to DNA in D. virilis is observed only at a distance of ~ 6.5 Kb from the TSS of Ranbp16 (Fig. 5). Within this range, a couple of transposon insertions are located. According to the data obtained by CAGE-seq (Cap Analysis Gene Expression), Ranbp16 gene in D. melanogaster has two TSS located at a distance of ~ 200 bp from each other, giving rise to slightly distinct transcripts in terms of the length of their 5’UTR. Transcripts with the shorter 5’UTR are transcribed throughout the development of Drosophila, whereas transcripts with the longer 5’UTR are present mostly at the larval stage (Fig. S4). Given that BEAF-32 is enriched within the upstream promoter of Ranbp16 in D. melanogaster, we assumed that in D. virilis the distance to the upstream promoter has been extended due to transposon insertions that may be spliced in the course of transcription. However, we failed to observe more than a single TSS by 5’RACE analysis at the larval and imago stages as well as in the gonads of D. virilis, suggesting either loss of the second promoter or its extremely low expression level.

Overall, these results indicate that the insulator protein BEAF-32 is enriched in the vicinity of the promoters of Myb and Ranbp16 genes embedded in distinct chromatin structures (euchromatic vs heterochromatic) in D. melanogaster and D. virilis, respectively.

Binding of BEAF-32 in the proximity of the transcription start sites is preserved in the course of evolutionary repositioning of gene loci between euchromatin and heterochromatin.
Considering the above results, two important questions may be asked. First, do the BEAF-32 binding sites in the proximity of promoter regions represent a peculiar property of Myb and Ranbp16 genes, or are they a common feature of heterochromatic genes? Second, do the BEAF-32 binding sites found in these promoters emerge in the course of adaptive evolution of genes transposed to heterochromatin? Alternatively, they might represent an ancestral feature which favors repositioning to the repressive environment without any deleterious impact on fitness.

To address these questions, we analyzed a representative set of more than 30 genes that reside in the pericentric heterochromatin of both arms of chromosome 2 in D. melanogaster, while in D. virilis the orthologs of these genes are clustered in different loci located in the euchromatin of the same chromosomal elements (Table S2) (21). Genes demonstrating the opposite scenario of repositioning were also included in the analysis. Among them we considered the heterochromatic genes Stim and Rrp47 that are located on the same scaffold_13050 as Myb and Ranbp16 of D. virilis, as well as the genes RpL15, Calr, Atg2 and CG40228 that are embedded in scaffold_12736 located near the chromocenter of D. virilis (Table S2) (18). In contrast to D. virilis, most of these genes with the exception of RpL15 and CG40228 are located in euchromatic regions of different chromosomal elements in D. melanogaster (Table S2).

Using these sets of genes, we performed enrichment analysis of BEAF-32, Pol II, H3K9me3 and ATAC-seq reads upstream and downstream of TSS of all these genes (Fig. 6). As indicated in Fig. 6A, the insulator protein BEAF-32 is enriched in the proximity of TSS or within 10 Kb upstream of TSS of both heterochromatic (Group 1) and euchromatic genes (Group 2) of D. melanogaster (Fig. 6A). The binding area of BEAF-32 is strongly correlated with the enrichment profile of Pol II and ATAC-seq reads (Fig. 6A). Importantly, analysis of D. virilis genes revealed a predominantly similar enrichment profile of BEAF-32 in both clusters of genes partitioned with regard to their heterochromatic or euchromatic location in the genome (Fig. 6B).
Taken together, these results indicate that BEAF-32 binding sites are present in the proximity of the promoters of genes that were repositioned into heterochromatin during the evolution of the *Drosophila* genus.

**Boundaries of heterochromatic gene loci are occupied by diverse insulator proteins in *D. melanogaster***

As demonstrated above, the insulator protein BEAF-32 is highly enriched in the proximity of promoters of heterochromatic genes. To further elucidate how widespread are insulators in heterochromatic regions, we analyzed the enrichment profiles of other *Drosophila* insulators, such as dCTCF, GAF, Pita, ZIPIC, Su(Hw) and Zw5, in the pericentric region of chromosome 2L of *D. melanogaster*. We find that, similar to BEAF-32, dCTCF and GAF have multiple binding sites, while Pita, ZIPIC and Su(Hw) are only modestly represented in the pericentric heterochromatin (Fig. S5A). Finally, Zw5 binding sites are completely absent from the studied region (Fig. S5A).

Next, we analyzed the enrichment profile of insulators upstream, downstream and within the bodies of heterochromatic genes located in the pericentric region of *D. melanogaster* chromosome 2L (Fig. S5B). We observed that BEAF-32 is not the only insulator found in the vicinity of TSS in heterochromatin genes. Although to a lesser extent, the insulators dCTCF and GAF are also enriched at the TSS of these genes (Fig. S5B). Interestingly, a few binding sites for Pita are present at the boundaries of gene loci as indicated by binding near the transcriptional start and transcriptional end sites (TES). Despite the binding of ZIPIC and Su(HW) to the pericentric DNA, the sites of their binding are not localized at the TSS and TES of the genes, suggesting that these insulators may be involved in long-range interactions of chromosomes, and hardly have any direct impact on gene expression (Fig. S5B).
Overall, these data suggest that the insulator proteins are involved in the maintenance of pericentric heterochromatin conformation and normal functioning of structural genes located in pericentromeric heterochromatin.

**Discussion**

During speciation, the genomes of *Drosophila* species underwent multiple chromosome rearrangements that disrupt gene order, modifying or preserving gene function (17, 42). In this study, we show that in the course of evolution of *Drosophila* species the essential *Myb* and *Ranbp16* genes have been repositioned into different chromatin types, i.e. euchromatin or heterochromatin. Despite the contrasting chromatin structure and local repressive environment of heterochromatic regions enriched with repetitive DNA, both genes were shown to be under purifying selection due to their highly conserved function. However, the overall gene organization may not be as conserved as the protein-coding region of heterochromatic genes. It was reported in a number of studies (18-21), that genes located in heterochromatic environment may increase their size by becoming targets for multiple TE insertions in their introns. Similar to the described evolutionary history of *RpL15* and *Yeti* (18, 19), the organization of *Myb* remained stable during evolution and includes a single short intron in the coding region of all species belonging to Sophophora and Drosophila subgenera. In contrast to *Myb*, the structure of *Ranbp16* gene embedded in heterochromatin has evolved to accumulate numerous TEs in its introns. These introns extended the gene length from ~40 kb in *D. mojavensis* to ~180 kb in *D. miranda*. The adaptive value of different intron lengths in the same gene in related species is unknown. A simple explanation of different intron content of the studied genes (*Myb* vs *Ranbp16*) stems from the fact that selective pressure favors short introns in highly expressed genes rather than in genes that are expressed at lower levels (43). In addition, a phenomenon called intron delay may limit the ability of cells with a short mitotic cycle to transcribe long
primary transcripts (44, 45). On this view, Myb, which is a highly expressed gene throughout Drosophila development, retained its organization regardless of its euchromatic or heterochromatic location. On the contrary, Ranbp16 gene, one of the members of importin-β family, is expressed at low level and probably exhibits tissue-specific patterns of expression. This may result in its more labile organization on the evolutionary timescale, with a remarkable increase in size in the heterochromatin context.

An unexplained question is the differential susceptibility of Ranbp16 introns to expansion. A dramatic increase of Ranbp16 gene length in several species occurred only due to the extension of certain but not all introns comprising the gene. It may be thought that selective pressure affects different parts of Ranbp16 gene to varying degrees, enabling expansion of some introns and preserving the length of others. In genomes of animals, plants, fungi, and protists, intron positions could be conserved throughout prolonged evolutionary times, suggesting their involvement in gene expression and regulation, including mRNA processing (46-48). Numerous studies in animals and plants have emphasized that introns may carry functional cis-acting elements affecting gene expression even in the absence of promoter sequences (49-53). Recently, it was shown that nucleosomes preferentially occupy exons rather than introns, and this phenomenon seems to be interconnected with specific histone modifications favoring gene expression (54-56). Thus, conservation of specific introns in the evolution of heterochromatic genes suggests their role in regulation of expression of these genes. However, the cross-talk between regulatory elements in introns, gene architecture, chromatin structure, and nucleosome positioning and modifications, especially in the heterochromatin, needs to be investigated in more detail.

According to studies of PEV, genes that are juxtaposed with heterochromatin by chromosomal rearrangements or transposition events exhibit a variegating phenotype resulting in silencing of genes due to heterochromatin environment (8). Given the peculiarities of heterochromatic genes, such as accumulation of TEs within their introns, one may suggest that
heterochromatic genes have evolved to utilize the advantage of the heterochromatic environment and became dependent on heterochromatin specific proteins, such as HP1α (4, 57-60). However, as was demonstrated previously, the regulatory regions including gene promoters apparently did not undergo drastic changes during their evolution in the heterochromatic environment (20). Moreover, evolutionary repositioning preferentially occurred only with gene clusters exhibiting close association with HP1α, suggesting that HP1α binding to these genes was present in the progenitor (21). Taken together, these observations allow one to suggest that certain gene loci are favored for repositioning from euchromatin to heterochromatin in the course of evolution. If this is the case, gene loci relocated to heterochromatin probably should retain the transcriptionally active euchromatin-like structure of chromatin capable of efficient transcription in the heterochromatin. Indeed, the proximal regulatory regions of heterochromatic genes are not occupied by the heterochromatic mark H3K9me3, forming a nucleosome-free binding platform for transcriptional factors and RNA polymerase II (Fig. 5 and 6).

How might these observations be explained? What determines the “immunity” of regulatory regions of heterochromatic genes to the repressive surroundings? In this study, we have examined the promoter of Myb gene in thirty Drosophila species and revealed common motifs between these sequences that serve as binding sites for the insulator protein BEAF-32 and transcriptional factor DREF. Originally, insulators were defined as regulators of interaction between enhancers and promoters able to block heterochromatin spreading and PEV (24, 25). DREF is known to mediate activation of transcription of genes involved in DNA replication and cell proliferation, including Myb gene (61, 62). Furthermore, cis-acting elements that exercise the transcriptional control of genes by DREF are conserved between such evolutionarily distant species as D. melanogaster and D. virilis (63). The DNA recognition motif for DREF (CGATA) is similar to that of BEAF-32 (TATCGATA), and the binding of DREF to DNA has been shown to antagonize the binding of BEAF-32 in vitro (64). Recent studies of the relationship between BEAF-32 and DREF showed that DREF co-localizes at the same genomic sites as BEAF-32 and
other insulator proteins and is enriched at the boundaries of topologically associated domains (TAD) (65). Together, these data suggest that DREF function is mediated by and probably depends on insulators on an evolutionary timescale.

A growing body of evidence suggests that insulators exercise diverse roles, including barrier function, and mediate short and long-distance chromosomal contacts at the genome-wide scale (66-70). Our enrichment analysis of ChIP-seq data indicated that the insulator protein BEAF-32 is enriched upstream of the TSS of heterochromatic genes in *D. melanogaster* and *D. virilis*, demarcating the euchromatin-heterochromatin border between the promoter and the surrounding heterochromatin (Fig. 6). Furthermore, using the same set of genes that reside in different types of chromatin in *D. melanogaster* and *D. virilis*, we show that BEAF-32 binding is preserved in the proximity of the promoter regions of these genes during evolution of different *Drosophila* species, suggesting that BEAF-32 binding is an ancestral property of these genes.

Along with BEAF-32, insulators dCTCF and GAF are also enriched at the TSS of heterochromatic genes, while the binding peaks for insulators Pita, ZIPIC and Su(Hw) are distributed across the pericentric region to a much lesser extent. The insulators dCTCF and GAF have many overlapping binding sites with BEAF-32 and DREF in the *Drosophila* genome (65). Furthermore, sites containing DREF without BEAF-32 are instead enriched with the Su(Hw) recognition sequences (65). Pita also belongs to a class of insulator proteins that preferentially bind to promoters near the TSS (71). ZIPIC is described as facilitator of long-distance chromosomal interactions (26, 71).

How this complex machinery works to facilitate continuous gene expression and chromosomal architecture remains an unresolved question. It is of note that a direct impact of insulator presence on gene expression has been established for the *D. melanogaster* GAGA factor (GAF) that resides in the *hsp70* promoter. GAF mediates the recruitment of chromatin remodeling factors, including SWI/SNF, the CHD, and the ISWI family complexes, that ensure formation of nucleosome-free region in the *hsp70* promoter (72-74). Studies of BEAF-32
demonstrated that most BEAF-associated genes are transcriptionally active or highly expressed and are associated with negative elongation factor NELF that stimulates transcription levels by inhibiting promoter-proximal nucleosome assembly (67, 75). This provides evidence that BEAF-32 facilitates high levels of gene expression.

Previously, it was shown that heterochromatic genes take advantage of repetitive surrounding and heterochromatin factors, such as HP1α, to facilitate their expression and probably long-distance interactions between enhancers and promoters (11, 22, 23, 76). On the other hand, an analysis conducted by Yasuhara et al (20) revealed that promoters of heterochromatic genes have not undergone major alterations after repositioning into the repetitive environment of heterochromatin rejecting the existence of heterochromatin-specific promoter. Along these lines, the observed binding of BEAF-32 in the proximity of gene promoters which underwent repositioning between euchromatin and heterochromatin in evolutionarily distant species of *Drosophila* is very intriguing. In cooperation with HP1α, the presence of BEAF-32 at gene promoters probably contributed to the formation of “proto-heterochromatic” gene loci in the ancestral species of *Drosophila* and thus facilitated their normal functioning in the heterochromatic environment. Recently, a remarkable peculiarity of HP1α binding at several gene loci has been described, whereby HP1α can be recruited to gene promoters independently of H3K9 methylation (77). Hence, one may hypothesize that BEAF-32 mediates the barrier function blocking the spreading of heterochromatin to the promoter regions of heterochromatic genes, while HP1α maintains a proper chromatin structure at and around such gene loci.

While it is clear that further studies are needed to elucidate all the factors required for normal gene functioning in the heterochromatic surroundings, our results provide an important insight into molecular mechanisms operating in the evolution of heterochromatic genes to provide their expression in the repressive environment.
Materials and methods

Drosophila genomes and sequence analyses

Drosophila genomes and gene sequences for comparative analysis were extracted from FlyBase and NCBI databases. Sequences of genomic regions containing Myb gene of virilis group species e.g. *D. lacicola, D. littoralis, D. borealis, D. flavomontana, D. lummei, D. ezoana* were fetched from unpublished data of Dr. Venera Tyukmaeva & Prof. Michael Ritchie from the University of St. Andrews, UK (personal communication). In the case of absence of gene annotation (e.g. for *D. guanche, D. subobscura*, most of the virilis group species), orthologs were retrieved with TblastN (78) using protein sequence of the most closely related species (i.e. *D. obscura* for *D. subobscura* and *D. guanche, D. virilis* for *D. lacicola* and other virilis species) as queries. All the query subjects mapped on the same DNA strand adjacent to each other with E-value > e-80 were considered as true and used for reconstruction of coding sequences. Sequences between mapped subjects were considered as introns. Putative transcriptional start sites (TSS) of poorly annotated genes were identified with blastN (79) using 1st exon sequence of related species. Blast results with E-value > e-60 and adjacent to annotated coding sequence at a distance not exceeding 600 nt were considered as true and the 1st mapped nucleotide as TSS. All essential information, including genes IDs, genomes IDs, and genomic coordinates of Myb and *Ranbp16* in all studied species, is listed in table S1. Orthologous sequences of Myb and *Ranbp16* of *Anopheles gambiae* were extracted from VectorBase (https://www.vectorbase.org/) by the numbers AGAP008160 – *Myb* and AGAP004535 – *Ranbp16*. Protein sequences of *Myb* and *Ranbp16* (also known as *Xpo7*) of mouse and human were extracted from UniProt (https://www.uniprot.org/). Protein motifs were scanned using the PROSITE database and methodology (80, 81).

Estimation of repeat content in intergenic regions and within studied genes was performed using RepeatMasker (82) and computationally predicted libraries of TEs generated
with REaS (83) that are available in FlyBase (ftp://ftp.flybase.net/genomes/aaa/transposable_elements/ReAS/v2/consensus_fasta/). For repeat masking of D. miranda genome, we used consensus sequences of TEs of D. pseudoobscura and D. persimilis, and for D. hydei we applied the library of D. mojavensis. TEs were classified using RepeatClassifier implemented in RepeatModeler software (84).

Multiple sequence alignment was performed with ClustalW (85) and Clustal Omega (86) programs (https://www.ebi.ac.uk/Tools/msa/) for nucleotide and amino acid alignments, respectively. Multiple protein alignments were visualized with Jalview (87).

Exon-intron structure of studied genes were visualized using WormWeb exon-intron graphic maker (http://wormweb.org/exonintron). Visualization of blast results was done with Kablammo (88).

Circular plot was made using Circos visualization tool (89). Flanking regions of Myb and Ranbp16 (20 Kb upstream and downstream from the gene location) were used instead of intergenic regions, due to the long distance between these genes in D. miranda (> 17 million bp) and low scaffold contiguity around these genes for D. persimilis and D. hydei.

ChIP-seq, RNA-seq and ATAC-seq analyses

Raw data of genome binding/occupancy (ChIP-seq), transcriptome (RNA-seq) and nucleosome (ATAC-seq) profiling were obtained from GEO database and used in the analyses. They include: GSE59965 - contains data for D. virilis including RNA-seq, ChIP-seq of H3K9me3 and RNA polymerase II performed using commercially available anti-H3K9me3 (ab8898, Abcam) and anti-RNA Pol II (ab5408, Abcam) antibodies; GSE35648 - contains data for both D. melanogaster and D. virilis including ChIP of BEAF-32 performed using antibodies generated against amino acids 1-83 of the major highly conserved isoform BEAF-32B in D. melanogaster (90); GSE43829 - contains RNA-seq as well as ChIP-seq of H3K9me3 and RNA polymerase II for D. melanogaster performed using aforementioned commercially available
antibodies ab8898 and ab5408; GSE56347 - includes ChIP-seq of HP1a for *D. melanogaster*
performed with polyclonal anti-HP1 (PRB-291C, Covance innovative); GSE102439 - includes
ATAC-seq data for *D. melanogaster and D. virilis*; finally, GSE85404, GSE33052, GSE70632
and GSE76997 - contains ChIP-seq data of dCTCF, Su(Hw), GAF, Pita, ZIPIC and Zw5 for *D. melanogaster*.

For analysis of sequence data, we used genome sequence and annotations released in FlyBase, *D. melanogaster* r.6.19 and *D. virilis* r1.06. Prior to mapping, all libraries were subjected to adapter clipping, filtering by length (>20 nt) and quality (80% of nt must have at least 20 Phred quality) using TrimGalore ([https://github.com/FelixKrueger/TrimGalore](https://github.com/FelixKrueger/TrimGalore)). Then, sequences were aligned to corresponding genomes using Bowtie (91) with the following settings: `-v 0 --best --strata`, retaining only aligned reads with zero mismatches. Output sequence alignment map (SAM) files were converted to binary (BAM) format using SAMtools (92). BAM files were converted to wiggle (WIG) format using bamToWig script ([https://github.com/craiglowe/bamToWig](https://github.com/craiglowe/bamToWig)). Aligned reads in wig format were visualized using the Integrative Genome Viewer (IGV) (93).

Peak calling was performed using MACS software (94) with the recommended parameters for narrow (PolII, BEAF-32, dCTCF, GAF, Su(HW), Pita, ZIPIC, Zw5) and broad peak calling (H3K9me3, HP1a). Enrichment analysis was performed using pipelines implemented in deepTools package (95) with the parameters including ignoring of duplicates.

For ATAC-seq analysis, reads that mapped on mitochondrial genomes were discarded, and peak calling was performed using Genrich ([https://github.com/jsh58/Genrich](https://github.com/jsh58/Genrich)) with the following settings: `--j -y -r -d 50`, including removal of PCR duplicates.

**Promoter analysis**

Because of insertion of DAIBAM MITE at the distance of 92 bp upstream from TSS of *Myb* in *D. virilis*, the promoter regions of *Myb* in studied species were shortened to 100 bp. After
sequence extraction, promoter regions of *Myb* and *Ranbp16* were searched for common motifs using MEME (96) and identification of matches to known transcription factors was performed by Tomtom (97) using OnTheFly (39) and REDfly v5.6 (40, 41) databases implemented in MEME Suite 5.0.5 (38).

**Sequence evolution and testing for selection**

Analysis of nucleotide substitutions per site was conducted in MEGA X (98) using the Tamura-Nei model (99). Rate variation among sites was modeled with a gamma distribution (shape parameter = 1). All positions containing gaps and missing data were eliminated (complete deletion option).

Ratio of nonsynonymous and synonymous substitutions (dN/dS) was estimated using PAL2NAL software (100) by converting multiple sequence alignment of proteins and the corresponding nucleotide sequences into a codon alignment, and the calculation of synonymous (dS) and non-synonymous (dN) substitution rates using codeml program implemented in PAML package (101).

**Cytology and DNA in situ hybridization**

*D. virilis* larvae were grown at 18°C on standard medium supplemented with live yeast solution for 2 days before dissection. Salivary glands from 3rd instar larvae were dissected in 45% acetic acid and squashed. DNA probes corresponding to *D. virilis Myb* (Dvir\'GJ18431; FBgn0205590) were prepared by PCR using gene-specific primers (Forward_591GCAAGTGCGAGCACTGAAAA; Reverse_TGCATACTGAGGTGTGCCAG). Then, DNA probe was biotinylated by nick translation using Biotin-14-dATP (Thermo Fisher Scientific, USA) as described in (102). Localization of the probe was made using the cytological map of *D. virilis* chromosomes (103). Images were obtained by binocular microscope Nikon Alphaphot-2 YS2 (Japan).
**RNA isolation, RT-PCR and 5’-RACE analysis**

Total RNA from 3rd instar larvae, adult females and gonads was isolated using Extract RNA reagent (Evrogen, Russia). Synthesis of the first strand of cDNA from total RNA and subsequent amplification of regions of interest were performed using MINT cDNA kit (Evrogen, Russia) following manufacturer’s instructions. For specific rapid amplification of cDNA 5’-end (5’-RACE) analysis, we applied two outward primers (primer1 5’-AGTAGTTGTGCGTAGCTGGA-3’; primer2 5’-GCTGCTTGCACAATGTTTCTA-3’) corresponding to the annotated 5’-fragment of *D. virilis Ranbp16* gene (Dvir\GJ18467; FlyBase ID: FBgn0205626). PCR reaction was conducted using Encyclo DNA polymerase (Evrogen, Russia). The resulting PCR fragments were cloned into pAL2-T vector (Evrogen, Russia) and sequenced using plasmid-specific primers. In all RT-PCR experiments, probes containing all components but lacking reverse transcriptase were used as negative controls. The obtained sequence of 5’UTR of *D. virilis Ranbp16* gene was deposited in GenBank under the number MN481598.

**ACCESSION NUMBERS**

The sequence of 5’UTR of *D. virilis Ranbp16* gene was deposited in GenBank under the number MN481598.

**Financial disclosure**

This work was supported by the Russian Foundation for Basic Research (grant number 19-04-00337). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Acknowledgements

We are grateful to Dr. Anton Golovnin from the Institute of gene biology RAS for ideas and critical comments on the manuscript. We thank Alexei M. Kulikov from the Koltsov Institute of developmental biology RAS for helpful advices and technical assistance. The bioinformatics was performed using the computational facilities of Engelhardt Institute of Molecular Biology RAS Genome center (http://www.eimb.ru/rus/ckp/ccu_genome_c.php).

References

1. Berger SL. The complex language of chromatin regulation during transcription. Nature. 2007;447(7143):407-12.

2. Eissenberg JC, Reuter G. Cellular mechanism for targeting heterochromatin formation in Drosophila. Int Rev Cell Mol Biol. 2009;273:1-47.

3. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell. 2010;143(2):212-24.

4. Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, et al. Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res. 2011;21(2):147-63.

5. Riddle NC, Jung YL, Gu T, Alekseyenko AA, Asker D, Gui H, et al. Enrichment of HP1a on Drosophila chromosome 4 genes creates an alternate chromatin structure critical for regulation in this heterochromatic domain. PLoS Genet. 2012;8(9):e1002954.

6. Pimpinelli S, Berloco M, Fanti L, Dimitri P, Bonaccorsi S, Marchetti E, et al. Transposable elements are stable structural components of Drosophila melanogaster heterochromatin. Proc Natl Acad Sci U S A. 1995;92(9):3804-8.
6487. Abramov YA, Shatskikh AS, Maksimenko OG, Bonaccorsi S, Gvozdev VA, Lavrov SA. The Differences Between Cis- and Trans-Gene Inactivation Caused by Heterochromatin in Drosophila. Genetics. 2016;202(1):93-106.

6518. Elgin SC, Reuter G. Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. Cold Spring Harb Perspect Biol. 2013;5(8):a017780.

6539. Sentmanat MF, Elgin SC. Ectopic assembly of heterochromatin in Drosophila melanogaster triggered by transposable elements. Proc Natl Acad Sci U S A. 2012;109(35):14104-9.

6550. Shatskikh AS, Abramov YA, Lavrov SA. Trans-inactivation: Repression in a wrong place. Fly (Austin). 2017;11(2):96-103.

6571. Wakimoto BT, Hearn MG. The effects of chromosome rearrangements on the expression of heterochromatic genes in chromosome 2L of Drosophila melanogaster. Genetics. 1990;125(1):141-54.

6591. Wheeler BS, Blau JA, Willard HF, Scott KC. The impact of local genome sequence on defining heterochromatin domains. PLoS Genet. 2009;5(4):e1000453.

6613. Coulthard AB, Alm C, Cealiac I, Sinclair DA, Honda BM, Rossi F, et al. Essential loci in centromeric heterochromatin of Drosophila melanogaster. I: the right arm of chromosome 2. Genetics. 2010;185(2):479-95.

6641. Dimitri P. Cytogenetic analysis of the second chromosome heterochromatin of Drosophila melanogaster. Genetics. 1991;127(3):553-64.

6661. Syrzycka M, Hallson G, Fitzpatrick KA, Kim I, Cotsworth S, Hollebakken RE, et al. Genetic and Molecular Analysis of Essential Genes in Centromeric Heterochromatin of the Left Arm of Chromosome 683 in Drosophila melanogaster. G3 (Bethesda). 2019;9(5):1581-95.

6691. Yasuhara JC, Wakimoto BT. Oxymoron no more: the expanding world of heterochromatic genes. Trends Genet. 2006;22(6):330-8.

6717. Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM. Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes. Genetics. 2008;179(3):1657-80.
Schulze SR, McAllister BF, Sinclair DA, Fitzpatrick KA, Marchetti M, Pimpinelli S, et al. Heterochromatc genes in Drosophila: a comparative analysis of two genes. Genetics. 2006;173(3):1433-67545.

Moschetti R, Celauro E, Cruciani F, Caizzi R, Dimitri P. On the evolution of Yeti, a Drosophila melanogaster heterochromatin gene. PLoS One. 2014;9(11):e113010.

Yasuha JC, DeCrease CH, Wakimoto BT. Evolution of heterochromatic genes of Drosophila. Proc Natl Acad Sci U S A. 2005;102(31):10958-63.

Caizzi R, Moschetti R, Piacentni L, Fanti L, Marsano RM, Dimitri P. Comparative Genomic Analyses Provide New Insights into the Evolutionary Dynamics of Heterochromatin in Drosophila. PLoS Genet. 2016;12(8):e1006212.

Eberl DF, Duyf BJ, Hilliker AJ. The role of heterochromatin in the expression of a heterochromatic gene, the rolled locus of Drosophila melanogaster. Genetics. 1993;134(1):277-92.

Yasuha JC, Wakimoto BT. Molecular landscape of modified histones in Drosophila heterochromatic genes and euchromatin-heterochromatin transition zones. PLoS Genet. 2008;4(1):e16.

Cho DH, Thienes CP, Mahoney SE, Analau E, Filippova GN, Tapscott SJ. Antisense transcription and heterochromatin at the DM1 CTG repeats are constrained by CTCF. Mol Cell. 2005;20(3):483-9.

Pikaart MJ, Recillas-Targa F, Felsenfeld G. Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators. Genes Dev. 1998;12(18):2852-62.

Maksimenko O, Bartkuhn M, Stakhov V, Herold M, Zolotarev N, Jox T, et al. Two new insulator proteins, Pita and ZIPIC, target CP190 to chromatin. Genome Res. 2015;25(1):89-99.

Phillips-Cremins JE, Corces VG. Chromatin insulators: linking genome organization to cellular function. Mol Cell. 2013;50(4):461-74.

Kyrchanova O, Georgiev P. Chromatin insulators and long-distance interactions in Drosophila. FEBS Lett. 2014;588(1):8-14.

Schwartz YB, Linder-Basso D, Kharchenko PV, Tolstorukov MY, Kim M, Li HB, et al. Nature and function of insulator protein binding sites in the Drosophila genome. Genome Res. 2012;22(11):2188-98.
30. Manak JR, Wen H, Van T, Andrejka L, Lipsick JS. Loss of Drosophila Myb interrupts the progression of chromosome condensation. Nat Cell Biol. 2007;9(5):581-7.

31. Okada M, Akimaru H, Hou DX, Takahashi T, Ishii S. Myb controls G(2)/M progression by inducing cyclin B expression in the Drosophila eye imaginal disc. EMBO J. 2002;21(4):675-84.

32. Drosophila 12 Genomes C, Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007;450(7167):203-18.

33. Gao JJ, Watabe HA, Aotsuka T, Pang JF, Zhang YP. Molecular phylogeny of the Drosophila obscura species group, with emphasis on the Old World species. BMC Evol Biol. 2007;7:87.

34. Gibbs AG, Matzkin LM. Evolution of water balance in the genus Drosophila. J Exp Biol. 2001;204(Pt 13):2331-8.

35. Jezovit JA, Levine JD, Schneider J. Phylogeny, environment and sexual communication across the Drosophila genus. J Exp Biol. 2017;220(Pt 1):42-52.

36. O'Grady PM, DeSalle R. Phylogeny of the Genus Drosophila. Genetics. 2018;209(1):1-25.

37. Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, et al. Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics. 2008;179(3):1601-55.

38. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37(Web Server issue):W202-8.

39. Shazman S, Lee H, Socol Y, Mann RS, Honig B. OnTheFly: a database of Drosophila melanogaster transcription factors and their binding sites. Nucleic Acids Res. 2014;42(Database issue):D167-71.

40. Bergman CM, Carlson JW, Celniker SE. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics. 2005;21(8):1747-9.

41. Rivera J, Kerenan SVE, Gallo SM, Hafon MS. REDfly: the transcriptional regulatory element database for Drosophila. Nucleic Acids Res. 2019;47(D1):D828-D34.

42. Ranz JM, Maurin D, Chan YS, von Grotthuss M, Hillier LW, Roote J, et al. Principles of genome evolution in the Drosophila melanogaster species group. PLoS Biol. 2007;5(6):e152.
72743. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. Selection for short introns in highly expressed genes. Nat Genet. 2002;31(4):415-8.

72944. Artieri CG, Fraser HB. Transcript length mediates developmental timing of gene expression across Drosophila. Mol Biol Evol. 2014;31(11):2879-89.

73145. Swinburne IA, Silver PA. Intron delays and transcriptional timing during development. Dev Cell.

732008;14(3):324-30.

73346. Carmel L, Rogozin IB, Wolf YI, Koonin EV. Patterns of intron gain and conservation in eukaryotic genes. BMC Evol Biol. 2007;7:192.

73547. Chorev M, Carmel L. The function of introns. Front Genet. 2012;3:55.

73648. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol.

7382003;13(17):1512-7.

73949. Chorev M, Joseph Bekker A, Goldberger J, Carmel L. Identification of introns harboring functional sequence elements through positional conservation. Sci Rep. 2017;7(1):4201.

74150. Gallegos JE, Rose AB. The enduring mystery of intron-mediated enhancement. Plant Sci.

7422015;237:8-15.

74351. Gallegos JE, Rose AB. Intron DNA Sequences Can Be More Important Than the Proximal Promoter in Determining the Site of Transcript Initiation. Plant Cell. 2017;29(4):843-53.

74552. Nott A, Meislin SH, Moore MJ. A quantitative analysis of intron effects on mammalian gene expression. RNA. 2003;9(5):607-17.

74753. Valencia P, Dias AP, Reed R. Splicing promotes rapid and efficient mRNA export in mammalian cells. Proc Natl Acad Sci U S A. 2008;105(9):3386-91.

74954. Dhani P, Saffrey P, Bruce AW, Dillon SC, Chiang K, Bonhoure N, et al. Complex exon-intron marking by histone modifications is not determined solely by nucleosome distribution. PLoS One.

7512010;5(8):e12339.

75255. Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nat Struct Mol Biol. 2009;16(9):990-5.
Venkatesh S, Workman JL. Histone exchange, chromatin structure and the regulation of transcription. Nat Rev Mol Cell Biol. 2015;16(3):178-89.

de Wit E, Greil F, van Steensel B. High-resolution mapping reveals links of HP1 with active and inactive chromatin components. PLoS Genet. 2007;3(3):e38.

Hearn MG, Hedrick A, Grigliat TA, Wakimoto BT. The effect of modifiers of position-effect variegation on the variegation of heterochromatic genes of Drosophila melanogaster. Genetics. 2011;188(2):651-62.

Lu BY, Emtage PC, Duyf BJ, Hilliker AJ, Eissenberg JC. Heterochromatin protein 1 is required for the normal expression of two heterochromatin genes in Drosophila. Genetics. 2000;155(2):699-708.

Lundberg LE, Stenberg P, Larsson J. HP1a, Su(var)3-9, SETDB1 and POF stimulate or repress gene expression depending on genomic position, gene length and expression pattern in Drosophila melanogaster. Nucleic Acids Res. 2013;41(8):4481-94.

Matsukage A, Hirose F, Yoo MA, Yamaguchi M. The DRE/DREF transcriptional regulatory system: a master key for cell proliferation. Biochim Biophys Acta. 2008;1779(2):81-9.

Sharkov NV, Ramsay G, Katzen AL. The DNA replication-related element-binding factor (DREF) is a transcriptional regulator of the Drosophila myb gene. Gene. 2002;297(1-2):209-19.

Kwon E, Seto H, Hirose F, Ohshima N, Takahashi Y, Nishida Y, et al. Transcription control of a gene for Drosophila transcription factor, DREF by DRE and cis-elements conserved between Drosophila melanogaster and virilis. Gene. 2003;309(2):101-16.

Hart CM, Cuvier O, Laemmli UK. Evidence for an antagonistic relationship between the boundary element-associated factor BEAF and the transcription factor DREF. Chromosoma. 1999;108(6):375-83.

Gurudatta BV, Yang J, Van Bortle K, Donlin-Asp PG, Corces VG. Dynamic changes in the genomic localization of DNA replication-related element binding factor during the cell cycle. Cell Cycle. 2013;12(10):1605-15.

Gomez-Diaz E, Corces VG. Architectural proteins: regulators of 3D genome organization in cell fate. Trends Cell Biol. 2014;24(11):703-11.
78067. Jiang N, Emberly E, Cuvier O, Hart CM. Genome-wide mapping of boundary element-associated
factor (BEAF) binding sites in Drosophila melanogaster links BEAF to transcription. Mol Cell Biol.
7822009;29(13):3556-68.

78368. Maksimenko O, Georgiev P. Mechanisms and proteins involved in long-distance interactions.
Front Genet. 2014;5:28.

78569. Ramirez F, Bhardwaj V, Arrigoni L, Lam KC, Gruning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9(1):189.

78770. Shrestha S, Oh DH, McKowen JK, Dassanayake M, Hart CM. 4C-seq characterization of Drosophila BEAF binding regions provides evidence for highly variable long-distance interactions between active chromatin. PLoS One. 2018;13(9):e0203843.

79071. Zolotarev N, Fedotova A, Kyrchanova O, Bonchuk A, Penin AA, Lando AS, et al. Architectural proteins Pita, Zw5, and ZIPIC contain homodimerization domain and support specific long-range interactions in Drosophila. Nucleic Acids Res. 2016;44(15):7228-41.

79372. Guertin MJ, Petesch SJ, Zobeck KL, Min IM, Lis JT. Drosophila heat shock system as a general model to investigate transcriptional regulation. Cold Spring Harb Symp Quant Biol. 2010;75:1-9.

79573. Lebedeva LA, Nabirochkina EN, Kurshakova MM, Robert F, Krasnov AN, Evgen’ev MB, et al. Occupancy of the Drosophila hsp70 promoter by a subset of basal transcription factors diminishes upon transcriptional activation. Proc Natl Acad Sci U S A. 2005;102(50):18087-92.

79874. Lomaev D, Mikhailova A, Erokhin M, Shaposhnikov AV, Moresco JJ, Blokhina T, et al. The GAGA factor regulatory network: Identification of GAGA factor associated proteins. PLoS One. 2002017;12(3):e0173602.

80175. Gilchrist DA, Nechaev S, Lee C, Ghosh SK, Collins JB, Li L, et al. NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. Genes Dev. 20032008;22(14):1921-33.

80476. Howe M, Dimitri P, Berloco M, Wakimoto BT. Cis-effects of heterochromatin on heterochromatic and euchromatic gene activity in Drosophila melanogaster. Genetics. 1995;140(3):1033-45.
80677. Figueiredo ML, Philip P, Stenberg P, Larsson J. HP1α recruitment to promoters is independent of H3K9 methylation in Drosophila melanogaster. PLoS Genet. 2012;8(11):e1003061.

80878. States DJ, Gish W. Combined use of sequence similarity and codon bias for coding region identification. J Comput Biol. 1994;1(1):39-50.

81079. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10.

81280. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, et al. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. 2002;3(3):265-74.

81481. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res. 2013;41(Database issue):D344-7.

81682. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. 2015.

81783. Li R, Ye J, Li S, Wang J, Han Y, Ye C, et al. ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLoS Comput Biol. 2005;1(4):e43.

81984. Smit AF, Hubley R. RepeatModeler Open-1.0. Available from http://www.repeatmasker.org.

82008. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673-80.

82486. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol. 2014;1079:105-16.

82687. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189-91.

82888. Wintersinger JA, Wasmuth JD. Kablammo: an interactive, web-based BLAST results visualizer. Bioinformatics. 2015;31(8):1305-6.

83089. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639-45.
Bushey AM, Ramos E, Corces VG. Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions. Genes Dev. 2009;23(11):1338-50.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMTools. Bioinformatics. 2009;25(16):2078-9.

Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178-92.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42(Web Server issue):W187-91.

Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2:28-36.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8(2):R24.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547-9.

Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993;10(3):512-26.

Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34(Web Server issue):W609-12.

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586-85591.

Lim J. In situ hybridization with biotinylated DNA. Dros Inf Serv. 1993;72:73-7.

Gubenko IS, Evgen'ev MB. Cytological and linkage maps of Drosophila virilis chromosomes. Genetica. 1984;65(2):127-39.
Figure legends
**Fig. 1.** Circular plot demonstrating the overall homology of genomic regions comprising *Myb* and *Ranbp16* genes in the species of *Drosophila* genus. Color nodes indicate homology of *Myb* (red nodes) and *Ranbp16* (blue nodes) and intergenic regions between these genes consisting of repeats (light blue nodes) and protein-coding genes (gray nodes) among the species of the *Sophophora* and *Drosophila* subgenera. Tracks of the plot indicate the region of comparison on chromosome X, coordinates of genes including *Myb* and *Ranbp16*, as well as the content of repeats in the plotted region. The phylogenetic tree indicates the relationships among species with estimated time of divergence according to Clark et al. (32), to Gao et al. (33) for the obscura group, O’Grady et al. (36) for the virilis group and Gibbs et al. (34) for the repleta group.

**Fig. 2.** Analysis of genomic regions comprising *Myb* and *Ranbp16* genes in *D. melanogaster* and *D. virilis*. A) Genomic map of whole assembled chromosome X of *D. melanogaster* with mapped ChIP-seq reads of heterochromatic markers H3K9me3 and HP1a and the region depicting *Myb* and *Ranbp16* gene location (studied genes are marked with bold font). B) DNA *in situ* hybridization of *Myb* gene probe to polytene chromosomes of *D. virilis*. Black arrow
indicates the hybridization signal at the base of chromosome X of *D. virilis*. C) Genomic map of scaffold_13050 of *D. virilis* and the proximal scaffold_12970 which is attributed to chromosome X of *D. virilis* according to Schaeffer et al. (37) with mapped ChIP-seq reads of H3K9me3 and the region containing *Myb* and *Ranbp16* genes (marked with bold font) on the larger scale.

Fig. 3. Comparative gene organization of *Myb* (A) and *Ranbp16* (B) genes among eight *Drosophila* species. White and black regions on the scheme denote untranslated and coding regions, respectively. The gray range between genes indicates homologous regions. Homology of introns is showed only for the *Myb* gene. Rightwards corner arrow at the beginning of genes indicates the transcriptional start site. Numbers beneath the introns reflect their size in base pairs. Introns that have been increased in size by transposable element insertion are highlighted by red bidirectional arrows. The 3’ untranslated regions (3’UTR) are not given for most of the genes due to the absence of their annotations and homology to closely related species. Bootstrap consensus phylogenetic tree is given according to Clark et al. (32), Gao et al. (33), O’Grady et al. (36) and Gibbs et al. (34) considering different groups of the *Drosophila* genus.
Fig. 4. Comparative promoter analysis of Myb gene in 30 Drosophila species. A) Distribution of common motifs and their orientation (shown by colored arrows) in the promoter of Myb. Prior to the analysis, orientation of all sequences was adjusted so that the transcription start site (TSS) would be on the right. B) Sequence logos of the common motifs among the studied Drosophila species with matches between motifs and binding sites of known transcription factors. Logos for binding sites of the factors are shown with the corresponding $p$- and $q$-values. The colors of motifs in A) and B) correspond to each other. Bootstrap consensus phylogenetic tree is given according to Clark et al. (32), Gao et al. (33), O’Grady et al. (36) and Jezovit et al. (35).
Fig. 5. Enrichment profiles of BEAF-32, RNA Pol II, H3K9me3, RNA-seq and ATAC-seq reads in the genomic regions comprising Myb and Ranbp16 genes in *D. melanogaster* and *D. virilis*. A) and B) Enrichment of ChiP-seq reads within Myb and Ranbp16 genic loci in *D. melanogaster*. C) The enrichment profile within the Myb and Ranbp16 genic loci and the intergenic regions between these genes in *D. virilis*. Mapped ChIP-seq reads without normalization (mapped reads), calculated areas of enrichment relative to the input data (peaks; *p*- and *q*-value < 0.05) are shown for BEAF-32, Pol II, H3K9me3 and ATAC-seq data. Summits of the enriched reads are shown for BEAF-32 (peaks summits). RNA-seq reads were normalized to the sequence depth (RPM, reads per million).
Fig. 6. Enrichment profiles and heatmaps of BEAF-32, RNA Pol II, H3K9me3 and ATAC-seq reads upstream and downstream of the transcriptional start site (TSS) of genes that were shown to be repositioned from euchromatin to heterochromatin during *Drosophila* evolution. A) and B) show the enrichment profiles of ChIP-seq reads of genic loci of *D. melanogaster* and *D. virilis*, respectively. Gene loci are subdivided into groups 1 and 2 corresponding to their heterochromatic (marked by green color) and euchromatic (red color) location. Note that the analyzed area for BEAF-32 enrichment includes 10 Kb upstream and 2 Kb downstream from TSS, while for RNA Pol II, H3K9me3 and ATAC-seq reads area includes 2 Kb both upstream and downstream from the TSS. Color codes below the heatmaps indicate fold enrichment.

Supporting Information Legends
Supplementary figure S1. The content of repetitive DNA in the genomic region of D. virilis comprising Myb and Ranbp16 genes. Classes and subclasses of transposable elements are indicated in different colors. Divergence of repeats is shown in the percentage of element coverage of the entire length of the consensus sequence where 100% should be considered as entire coverage of repeat consensus sequence on genome region.

Supplementary figure S2. The estimated evolutionary divergence (A and B) and the ratio of non-synonymous and synonymous substitutions (C and D) across Diptera species. A and B) The number of base substitutions per site (Distance) for Myb and Ranbp16 genes, respectively. Standard error estimate(s) are shown above the diagonal (Standard error). Analyses were conducted using the Tamura-Nei model. All positions containing gaps were discarded. C and D) Results of dN/dS test for Myb and Ranbp16 genes, respectively. Estimation was performed in relation to D. melanogaster sequence.

Supplementary figure S3. Multiple protein alignment of Myb type HTH DNA-binding domain of Myb (A) and Importin-β domain of Ranbp16 (B) across selected taxa.

Supplementary figure S4. The intergenic region between Ranbp16 gene and the adjacent Stim gene of D. melanogaster exported from UCSC genome browser. CAGE-seq data indicate the beginning of 5’UTR (5’untranslated region) on each DNA strand. Top upper track shows the promoter region coordinates fetched from Eukaryotic Promoter Database (EPD) (https://epd.epfl.ch/index.php). Sequence conservation is shown for 27 insects presented in the UCSC database.

Supplementary figure S5. The enrichment profile of insulator proteins in the heterochromatin of D. melanogaster. A) The enrichment profile of insulators BEAF-32, dCTCF, GAF, Pita, ZIPIC, Su(Hw), Zw5 and heterochromatic marks H3K9me3 and HP1a in the pericentric region of chromosome 2L of D. melanogaster. The euchromatin-heterochromatin border is depicted by dashed line according to the mapping profile of H3K9me3 and HP1a. Mapped ChIP-seq reads without normalization (mapped reads) and calculated areas of enrichment relative to the input data with the p- and q-values < 0.05 (peaks) are shown for each factor. RNA-seq reads were normalized to the sequence depth (RPM, reads per million). B) The enrichment profile of insulator proteins and H3K9me3 upstream, within and downstream of heterochromatic genes located in pericentric region of chromosome 2L of D. melanogaster.
Supplementary table S1. List of the Myb and Ranbp16 orthologs and their coordinates in the genome.

Supplementary table S2. List of the studied heterochromatic genes of D. melanogaster and their orthologs in D. virilis with coordinates in the genome.