Enriched G-quadruplexes on the *Drosophila* Male X Chromosome Function as Insulators of Dosage Compensation Complex

Authors: Sheng-Hu Qian, Lu Chen, Zhen-Xia Chen*

Affiliations:

Hubei Key Laboratory of Agricultural Bioinformatics, College of Life Science and Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, PR China

*Correspondence to: zhen-xia.chen@mail.hzau.edu.cn

Abstract: The evolution of sex chromosomes has resulted in half X chromosome dosage in males as females. Dosage compensation, or the two-fold upregulation in males, was thus evolved to balance the gene expression between sexes. However, the step-wise evolutionary trajectory of dosage compensation during Y chromosome degeneration is still unclear. Here, we show that the specific structured elements G-quadruplexes (G4s) are enriched on the X chromosome in *Drosophila melanogaster*. Meanwhile, on the X chromosome, the G4s are underrepresented on the H4K16 acetylated regions and the binding sites of dosage compensation complex male-specific lethal (MSL) complex. Peaks of G4 density and potential are observed at the flanking regions of MSL binding sites, suggesting G4s act as insulators to precisely up-regulate certain regions in males. Thus, G4s may be involved in the evolution of dosage compensation process through fine-tuning one-dose proto-X chromosome regions around MSL binding sites during the gradual Y chromosome degeneration.

One Sentence Summary: G-quadruplexes act as insulators to precisely up-regulate X chromosome in males.

Main Text:

Dosage compensation in flies represents a prime example of fine-tuning gene expression of a whole chromosome. It balanced the gene expression between females (XX) and males (XY) through the twofold upregulation of X chromosome in males mediated by male-specific lethal (MSL) complex (*1*, *2*). MSL complex lands on genomic regions called High-Affinity Sites (HAS), and then spreads to acetylate histone H4 lysine 16 (H4K16ac) across the entire X chromosome (*3*). Recent advances reveal that the complex interplay among the epigenetic landscape, transcription, and chromosome conformation achieves male-specific X chromosome upregulation (*4*, *5*). However, the detailed mechanisms underlying dosage compensation are still obscure.

G-quadruplexes (G4s), the four-stranded helical structures formed by guanine-rich DNA or RNA (Fig. 1A), impose a substantial mutational burden on the genome (*6-8*). However, a growing body of evidence implicating G4s as regulators of gene expression (*9, 10*). To explore the functionality of G4s, we systematically searched for potential G4s (pG4s) in 20 *Drosophila* species, and identified 43,244-120,764 pG4s across their genomes (Fig. S1). Specifically, 46,435 pG4s were found in *D. melanogaster* (1.12% of the genome), including 15,663 pG4s in intergenic regions, and even 8.5 times as high as in randomly shuffled genome (Fig. S2). The density of pG4s in intergenic regions was higher than that in genic regions (Fig. 1B). Besides, although GC contents of sliding windows across the genome were positively correlated with the pG4 number (Fig. S3; $r = 0.359$, $P < 2.2 \times 10^{-16}$), the GC contents of intergenic regions were lower than genic regions (Fig. S4), suggesting a maintenance of intergenic pG4s as functional elements under selection. Alternatively, pG4s may be neutral or nearly neutral, and more pG4 were maintained in intergenic regions merely due to less selective constraints.

Considering repetitive regions encounter less selective constraints and nonfunctional elements were thus likely to accumulate there instead of non-repetitive regions, we classified intergenic pG4s into repetitive pG4s and non-repetitive pG4s based on whether or not they were overlapped with repetitive regions. We found that intergenic non-repetitive regions had the highest density with an average of 0.42/kb, followed by genic regions with an average of 0.32/kb and intergenic repetitive regions with an average of 0.23/kb (Table S1). This result suggested that intergenic pG4s, at least part of them, could be maintained as functional elements.

The hemizygosity of X chromosome in males led to different evolutionary forces acting on autosomes and X chromosome, followed by different distribution patterns of non-neutral elements in the genomes. We thus investigated the chromosomal distribution of pG4s to better understand why the genome, especially for intergenic regions, encodes so many of them. We found that pG4s were enriched on the X chromosome (2,552, 21.4 %) (Fig. 2A; P < 2.2 × 10-16, Fisher's exact test). Such enrichment still existed even after the normalization for GC contents (Fig. S5). Interestingly, the pG4s on the X chromosome were longer (Fig. 2B; P < 2.2 × 10-16, Kruskal-Wallis test) and had higher quadruplex propensity than autosomal pG4s (Fig. 2C; $P < 3.965 \times 10^{-13}$, Kruskal-Wallis test), suggesting that pG4s, especially X-linked pG4s, were not neutral.

The distribution of sex-biased elements was shaped by various contrasting forces they were subjected to (*11, 12*). To find out some clues on the evolutionary mechanisms of pG4s based on the distribution of sex-biased pG4s, we classified all expressed pG4s into three groups, including male-biased, female-biased and unbiased pG4s. Among the expressed pG4s, 474 (7.56%) were male-biased expressed, 727 (11.60%) were female-biased expressed and 5,065 (80.83%) were unbiased. Compared with male-biased pG4s, female-biased pG4s had greater expression level (Fig. S6; $P = 3.269 \times 10^{-5}$, Wilcoxon test), quadruplex structural scores (Fig. S7; $P = 1.823 \times 10^{-5}$, Wilcoxon test) and length (Fig. S8; $P = 1.675 \times 10^{-5}$, Wilcoxon test), suggesting that they were maintained under selection.

We then investigated the distribution of these expressed pG4s across the chromosomes, and found that male-biased pG4s with expression were underrepresented on the X chromosome ($P < 0.05$, Wilcoxon), while female-biased pG4s were overrepresented on the X chromosome (Fig. 2D; $P < 2.2 \times 10^{-16}$, Wilcoxon test), consistent with the pattern of protein-coding genes (*13, 14*) and non-coding RNAs (*15*). The concurrent underrepresentation of X-linked male-biased pG4s and overrepresentation of female-biased pG4s could be explained by sexual antagonism for dominant genes (*16*) or the repression of dosage compensation (*17*) by G4s.

G4s were reported to suppress gene expression (*18*, *19*). To test whether the enrichment of intergenic pG4s was responsible for the male-specific fine-tuning of X chromosome in dosage compensation, we then took advantage of RNA-seq data from fly tissues to measure the expression of pG4s. The pG4s with higher expression than 30% genic pG4s in any tissue were taken as expressed pG4s, while others were taken as unexpressed pG4s. In total, 21,009 of pG4s were expressed, including 2,739 (13.04%) intergenic pG4s. Expressed pG4s were underrepresented in intergenic regions. Moreover, we found lower expression level of intergenic pG4s than genic pG4s (Fig. S9; $P < 2 \times 10^{-16}$, Wilcoxon test), supporting that pG4s may suppress gene expression. We further tested whether the density of pG4s in a region was associated with its expression level with a sliding window strategy, and found windows with higher pG4 density had lower expression than those with lower pG4 density (Fig. 3A; $P < 6.14 \times 10^{-14}$, Wilcoxon test). The negative correlation between pG4 density and expression levels suggested the role of pG4s as suppressors. Next, we performed RNA-seq of *Drosophila* S2 cells treated with TMPyP4, a type of G-quadruplex stabilizer (Fig. 3B), and found the expression level of pG4s decreased with the increase of TMPyP4 concentration, supporting again the role of pG4s as suppressors.

The pG4s analyzed above were computationally predicted. To validate our observations, we thus reanalyzed the patterns of G4s identified with G4-seq in *Drosophila melanogaster (20)*. In total, 14,946 (32.5%, with K+ as stabilizing agent of G4-seq) and 17,495 pG4s (38.1%, with K+ and PDS as stabilizing agent of G4-seq) were supported by G4-seq (Table S2). The G4s identified by G4-seq on the X chromosome had higher density (~1.9 fold), more stable structures, and greater length than those on the autosomes, consistent with observations from pG4s predicted in silico (Fig. S10-S12).

We further explored whether the enrichment of X-linked pG4s with decreased expression was associated with X-specific upregulation in males. We found the pG4 density at HAS (Genomic coordinates from (*3*)) on the X chromosome (0.263/kb) was not only lower than the average pG4 density on the X chromosome (0.488/kb), but also than that across the whole genome (0.335/kb). The pG4 density in H4K16ac regions on the X chromosome (0.357/kb) was lower than the average pG4 density on the X chromosome, although higher than that in H4K16ac regions of the whole genome (0.357/kb) (Fig. 3C). We further calculated the pG4 density in H4K16ac regions and their flanking regions on autosomes and the X chromosome in both sexes, and found highest pG4 density in the H4K16ac flanking regions on the male X chromosome (Fig. 3D). Peaks of both the density (Fig. 3E and Fig. S13) and the quadruplex propensity of pG4s (Fig. 3F and Fig. S14) were found at the boundaries of H4K16ac regions, suggesting that pG4s confined the acetylated regions mediated by MSL complex in dosage compensation.

To explore the evolutionary patterns of pG4, we further dated the pG4s based on their presence or absence in the phylogenetic tree of 20 *Drosophila* species (*21, 22*) (Fig. 4A). We found 3,328 (26.4%) pG4s were specific to *D.melanogaster*, and the correlation between age and pG4 number was -0.57 (Spearman correlation test). Moreover, the younger pG4s showed an excess on the X chromosome, while the older pG4s showed a paucity (Fig. S15), suggesting faster birth and death of X-linked pG4s than autosomal pG4s. We then separated pG4s into three groups (young, middle and old) based on their age, and observed that younger pG4s on the X chromosome were more likely to be female-biased (Fig 4B; $P < 0.05$, Fisher), suggesting that newly formed pG4s were more likely to repress dosage compensation. As a validation, we calculated the density of pG4s in *Drosophila miranda* and *Drosophila pseudoobscura* with a series of X chromosomes of diverse ages. The XL chromosomes of ~60 million years old in their genomes were homologous to *D. melanogaster* X, and the XR chromosomes originated 15 million years ago, while the neo-X chromosome becoming a sex chromosome only 1 million years ago was specific to *D. miranda* (*23-25*). As expected, our result showed that the pG4 density on these sex chromosomes was higher than autosomes, and the pG4 density on the younger XR and neo-X were higher than that on the older XL (Fig. 4CD).

A model was thus proposed for the evolution of G4s on the X chromosome (Fig. 4E). In the gradual degradation of Y chromosome, dosage compensation was required for one-dose proto-X-linked genes. To fine-tune the expression of these one-dose genes locally instead of chromosome-wide, G4s formed flanking a newly evolved HAS and acted as insulators to protect two-dose proto-X-linked regions from acetylation and upregulation. With the extension of one-dose proto-X regions, G4s were selected against to allow their upregulation.

Acknowledgments: We thank Huijuan You, Dengguo Wei, Wenqiang Wu for providing TMPyP4, Zhihui Zhu for the gift of S2 cell. We also thank Manyuan Long, Brian Oliver, Jenny Graves, Charles Vinson, Xiao-Qin Xia, Yikang S. Rong, Qi Zhou, Jian Lu, Xionglei He, Jian-Quan Ni, Yu-Jie Fan and Wensheng Wei for helpful comments. Finally, I want to show my highest blessing to my father and mother, without them, there would be no me.

Funding: Z.C. was supported by Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (Program No.2016RC011).

Author contributions:

S. Q. and Z.C. designed the project. S. Q. performed the analysis. L. C. carried out the experiment. S. Q. and Z. C. wrote the manuscript with all authors contributing to writing.

C**ompeting interests:** Authors declare no competing interests.

Data and materials availability:

Data for analyzing G-qudruplex expression alteration under different TMPyP4 concentration from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE131691.

Supplementary Materials:

Materials and Methods

Figures S1-S15

Materials and Methods

Genomes

Genome sequences of *Drosophila menalogaster* (dmel_r6.19_FB2017_06), *Drosophila pseudoobscura/* (Dpse_r3.04) were downloaded from flybase (http://flybase.org/). 18 genome sequences of D. *ana* (dana_caf1), D. *bip* (Dbip_2.0), D. *ere* (dere_caf1), D. *fic* (Dfic_2.0), D. *kik* (Dkik_2.0), D. *moj* (dmoj_caf1), D. *sec* (dsec_caf1), D. *tak* (Dtak_2.0), D. *wil* (dwil_caf1), D. *bia* (Dbia_2.0), D. *ele* (Dele_2.0), D. *eug* (Deug_2.0), D. *gri* (dgri_caf1), D. *per* (dper_caf1), D. *rho* (Drho_2.0), D. *sim* (ASM75419v2), D. *vir* (dvir_caf1), D. *yak* (dyak_caf1) were downloaded from National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/).

Species	**Genome source**	**Genome version**
Drosophila ananassae	https://www.ncbi.nlm.nih.gov/genome	dana_caf1
Drosophila bipectinata	https://www.ncbi.nlm.nih.gov/genome	Dbip_2.0
Drosophila erecta	https://www.ncbi.nlm.nih.gov/genome	dere_caf1
Drosophila ficusphila	https://www.ncbi.nlm.nih.gov/genome	Dfic_2.0
Drosophila kikkawai	https://www.ncbi.nlm.nih.gov/genome	Dkik_2.0
Drosophila mojavensis	https://www.ncbi.nlm.nih.gov/genome	dmoj_caf1
Drosophila sechellia	https://www.ncbi.nlm.nih.gov/genome	dsec_caf1
Drosophila takahashii	https://www.ncbi.nlm.nih.gov/genome	Dtak_2.0
Drosophila willistoni	https://www.ncbi.nlm.nih.gov/genome	dwil_caf1
Drosophila biarmipes	https://www.ncbi.nlm.nih.gov/genome	Dbia_2.0
Drosophila elegans	https://www.ncbi.nlm.nih.gov/genome	Dele_2.0
Drosophila eugracilis	https://www.ncbi.nlm.nih.gov/genome	Deug_2.0
Drosophila grimshawi	https://www.ncbi.nlm.nih.gov/genome	dgri_caf1
Drosophila persimilis	https://www.ncbi.nlm.nih.gov/genome	dper_caf1
Drosophila rhopaloa	https://www.ncbi.nlm.nih.gov/genome	Drho_2.0

Drosophila simulans	https://www.ncbi.nlm.nih.gov/genome	ASM75419v2
Drosophila virilis	https://www.ncbi.nlm.nih.gov/genome	dvir_caf1
Drosophila yakuba	https://www.ncbi.nlm.nih.gov/genome	dyak_caf1
Drosophila miranda	https://www.ncbi.nlm.nih.gov/genome	DroMir_2.2
Drosophila pseudoobscura pseudoobscura	http://flybase.org/	Dpse_r3.04
Drosophila melanogaster	http://flybase.org/	dmel_r6.19

Detecting pG4s in the fly genomes and simulated genome

We used R package G4Hunter to identify pG4s in Drosophila melanogaster genome (26). G4Hunter predicted pG4s using a score function in which positive scores were given for Guanine while negative scores were given for Cytosine in each strand of DNA sequences. To avoid overestimation of G4 number, overlapping pG4 sequences were fused. We specified cutoff as 1.50 to find pG4s with scores no less than 1.50. Meanwhile, we applied a sliding window strategy (window size: 5kb, step size: 2.5kb) to assess the correlation between number of pG4s and GC content. We searched pG4s on chromosome 2L, 2R, 3L, 3R, 4, X, and Y, and identified 45,943 pG4s in total (referred to as whole-genome pG4s).

Then we generated 1,000 artificial randomized fly genomes to assess statistical significance of our data. We randomly shuffled the genome while maintained their original DNA composition and length by shuffleseq, implemented in EMBOSS (v6.5.7) (32). Therefore, we compared the number of pG4s in simulated genome with that of in original genome but the identical nucleotide compositions and lengths.

The classification of pG4s

We classified pG4s into repetitive and nonrepetitive PG4s based on repeats annotations produced by RepeatMasker (http://www.repeatmasker.org; last accessed Feb. 2, 2018) from UCSC Genome Browser. RepeatMasker classified low complexity regions and repeat regions into LTR, simple repeats, long Interspersed Elements (LINE), short Interspersed Elements (SINE) and other types of repeates. We referred the G4s having at least 1 nt overlap with any repeat as repetitive G4s and others as nonrepetitive G4s.

We also classified pG4s into genic and intergenic PG4s based on General Transfer Format (GTF) format gene annotation of D. melanogaster (Flybase r6.19) downloaded from Flybase (http://flybase.org/; last access Jan 27, 2018). We used the software bedtools command intersect (27) (v2.25.0) to search pG4s which had at least 1 nt overlap with any protein-coding genes, non-coding genes or psuedogenes and classified them into genic pG4s. Other pG4s were classified into intergenic pG4s. Furthermore, we separated pG4s into five groups based on their different length. The 20, 40, 60 and 80 quantiles of their length were 25, 28, 34 and 41. Thus pG4s were classified into five groups: SS, S, M, L and LL. GC contents in every region were calculated by bedtools nuc.

Cell culture and treatment

Drosophila line 2 cells (S2) were maintained at 28 □ in Hyclone TNM-FH insect medium containing 10% Gibico serum and antibiotics (0.5 U/ml penicillin and 0.5 µg/ml streptomycin). Cells were treated with different concentration of TMPyP4 (25umol/L, 50umol/L, 100umol/L, respecticely), and DMSO (100µg/L) was severed as control. After 48 hours, the S2 cells were sampled for extraction of RNA.

RNA extraction and sequencing

RNA samples from each individual were exacted using Trizol (from Invitrogen) according to the manufacturer's protocol. Stranded PolyA+ RNA libraries were prepared with in-house kits (Majorbio). The library quality was assessed by checking the distribution of the fragment size using Agilent 2100 bioanalyzer (Agilent) and the quantity of the libraries were measured using qRT-PCR (TaqMan Probe). In total, 12 qualified strand-specific cDNA libraries were constructed and were sequenced on the Illumina HiSeq System.

Expression profiling

RNA-seq data from fly somatic and gonadal tissues were downloaded from Gene Expression Omnibus (GEO) in NCBI (http://www.ncbi.nlm.nih.gov/geo/; accession numbers: GSE44612) (28). We then aligned sequence data, as well as raw reads in FASTQ format, by HISAT2 (v2.0.4) (29) to the reference genome (-x) and annotation (--known-splicesite-infile). To estimate the expression level of pG4s, we quantified counts of reads overlapping flanking 1 kb of each pG4 by slop and coverage implemented in bedtools, and normalized the read counts by R package DESeq2 (v1.18.1) (30). For each pG4, maximum normalized read counts among different tissues were regarded as expression level of pG4.

To define expression status of intergenic pG4s with high confidence, we used a strict cutoff derived from the expression level of genic pG4s. The median expression level of all genic pG4s was taken as the expression cutoff for intergenic pG4s. Then intergenic pG4s with normalized read count higher than the cutoff were defined as expressed pG4s, and others were unexpressed.

Dating G-quadruplexes on Phylogenetic Tree of Drosophila

To date all pG4s on Drosophila phylogenetic tree, we applied a modified pipeline based on previous method (21, 22). We performed whole-genome pairwise alignments between D. melanogaster and other 19 Drosophila species (D.ananassae, D.biarmipes, D.bipectinata, D.elegans, D.eugracilis, D.ficusphila, D.kikkawai, D.mojavensis, D.pseudoobscura, D.rhopaloa, D.simulans, D.takahashii, D.virilis, D.yakuba, D.erecta, D.sechellia, D.willistoni, D.grimshawi and D.persimilis). Based on UCSC genome browser mannual, we generated reciprocal best LiftOver chain files from the alignments and aligned D. melanogaster to other Drosophila species. We defined a pG4 as present in a species if it could be projected to the species no matter whether it was still a pG4s motif in the species. Finally, we dated all pG4s on the phylogenetic tree and assigned pG4s into different evolutionary branches. The pG4s assigned to branch1-3 were classified as young, those in branch4-6 were classified as middle and the rest were regarded as old.

MSL ChIP-seq and HAS analysis

ChIP-seq data for H4K16ac analysis were downloaded from NCBI under accession number GSE109901. We aligned the sequence data by bowtie2 (31) to the reference genome with default parameter. BAM files from ChIP-seq were used for peak calling with MACS2 (32). A bed file containing genomic coordinates of HAS was downloaded from (3) and converted to corresponding version by UCSC liftover with default parameter.

Profile of G4 in H4K16ac (scale to 1 kb) and flanking regions (500 bp) was analyzed by deeptools.

References and Notes:

1. W. Jordan, 3rd, L. E. Rieder, E. Larschan, Diverse Genome Topologies Characterize Dosage Compensation across Species. *Trends Genet* **35**, 308-315 (2019).
2. J. C. Lucchesi, M. I. Kuroda, Dosage compensation in Drosophila. *Cold Spring Harb Perspect Biol* **7**, (2015).
3. F. Ramirez *et al.*, High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in Drosophila. *Mol Cell* **60**, 146-162 (2015).
4. T. Conrad, A. Akhtar, Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nat Rev Genet* **13**, 123-134 (2012).
5. M. Samata, A. Akhtar, Dosage Compensation of the X Chromosome: A Complex Epigenetic Assignment Involving Chromatin Regulators and Long Noncoding RNAs. *Annu Rev Biochem* **87**, 323-350 (2018).
6. A. B. Sahakyan, P. Murat, C. Mayer, S. Balasubramanian, G-quadruplex structures within the 3' UTR of LINE-1 elements stimulate retrotransposition. *Nat Struct Mol Biol* **24**, 243-247 (2017).
7. D. Rhodes, H. J. Lipps, G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**, 8627-8637 (2015).
8. C. Ribeyre *et al.*, The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS Genet* **5**, e1000475 (2009).
9. A. Siddiqui-Jain, C. L. Grand, D. J. Bearss, L. H. Hurley, Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A* **99**, 11593-11598 (2002).
10. R. Hansel-Hertsch *et al.*, G-quadruplex structures mark human regulatory chromatin. *Nat Genet* **48**, 1267-1272 (2016).

11. R. P. Meisel, J. H. Malone, A. G. Clark, Disentangling the relationship between sex-biased gene expression and X-linkage. *Genome Res* **22**, 1255-1265 (2012).
12. S. L. Allen, R. Bonduriansky, S. F. Chenoweth, The genomic distribution of sex-biased genes in drosophila serrata: X chromosome demasculinization, feminization, and hyperexpression in both sexes. *Genome Biol Evol* **5**, 1986-1994 (2013).
13. M. Parisi *et al.*, Paucity of genes on the Drosophila X chromosome showing male-biased expression. *Science* **299**, 697-700 (2003).
14. J. M. Ranz, C. I. Castillo-Davis, C. D. Meiklejohn, D. L. Hartl, Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science* **300**, 1742-1745 (2003).
15. G. Gao *et al.*, A long-term demasculinization of X-linked intergenic noncoding RNAs in Drosophila melanogaster. *Genome Res* **24**, 629-638 (2014).
16. W. R. Rice, Sex Chromosomes and the Evolution of Sexual Dimorphism. *Evolution* **38**, 735-742 (1984).
17. H. Ellegren, J. Parsch, The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet* **8**, 689-698 (2007).
18. D. Dutta *et al.*, Cell penetrating thiazole peptides inhibit c-MYC expression via site-specific targeting of c-MYC G-quadruplex. *Nucleic Acids Res* **46**, 5355-5365 (2018).
19. S. Ravichandran *et al.*, Genome-wide analysis of regulatory G-quadruplexes affecting gene expression in human cytomegalovirus. *PLoS Pathog* **14**, e1007334 (2018).
20. G. Marsico *et al.*, Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res*, (2019).
21. Y. E. Zhang, M. D. Vibranovski, B. H. Krinsky, M. Long, Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Res* **20**, 1526-1533 (2010).
22. Z. X. Chen, B. Oliver, Y. E. Zhang, G. Gao, M. Long, Expressed Structurally-stable Inverted Duplicates in Mammalian Genomes as Functional Noncoding Elements. *Genome Biol Evol*, (2017).
23. C. E. Ellison, D. Bachtrog, Dosage compensation via transposable element mediated rewiring of a regulatory network. *Science* **342**, 846-850 (2013).
24. D. Bachtrog, B. Charlesworth, Reduced adaptation of a non-recombining neo-Y chromosome. *Nature* **416**, 323-326 (2002).
25. A. B. Carvalho, A. G. Clark, Y chromosome of D. pseudoobscura is not homologous to the ancestral Drosophila Y. *Science* **307**, 108-110 (2005).
26. V. Brazda *et al.*, G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics*, (2019).
27. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
28. Z. X. Chen *et al.*, Comparative validation of the D. melanogaster modENCODE transcriptome annotation. *Genome Res* **24**, 1209-1223 (2014).

29. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).

30. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

31. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

32. Y. Zhang *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

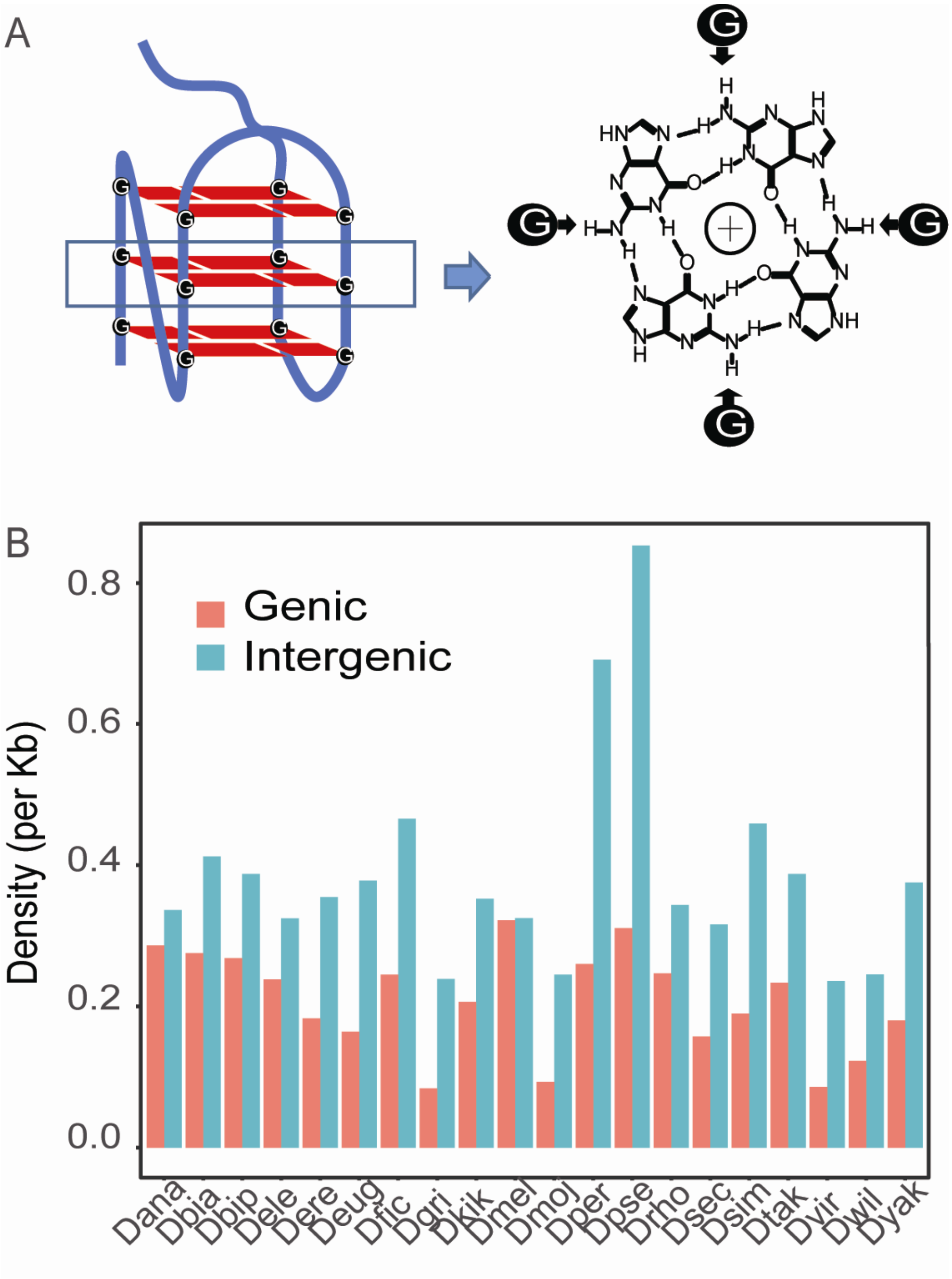

Fig. 1. The enrichment of pGs in intergenic regions. (A) The structure of G-quadruplex. (B) The distribution of pG4s in the genome of 20 fly species. The G4s that have any overlap with gene region were referred as genic G4s, others were assigned as intergenic ones.

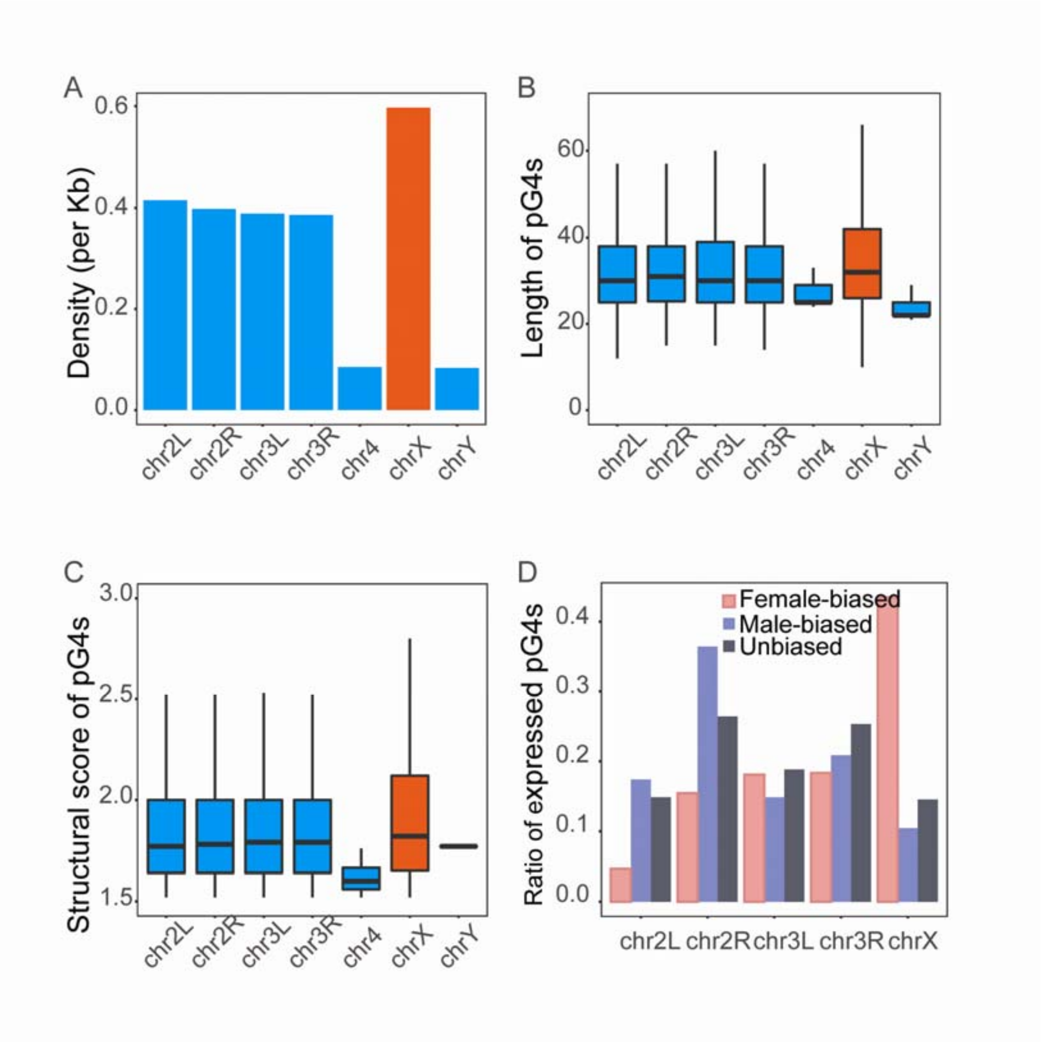

Fig. 2. The best performance of X-linked pG4s in chromosomal distribution, length and structural stability relative to others. (A) The X chromosome has the highest density of pG4s than other chromosomes. (B) X-linked pG4s are longest than others. (C) The pG4s on the X chromosome are the most structurally stable than others. (D) The pG4s with male-biased expression were enriched on the X chromosome, while those with female-biased expression showed a deficiency on the X chromosome.

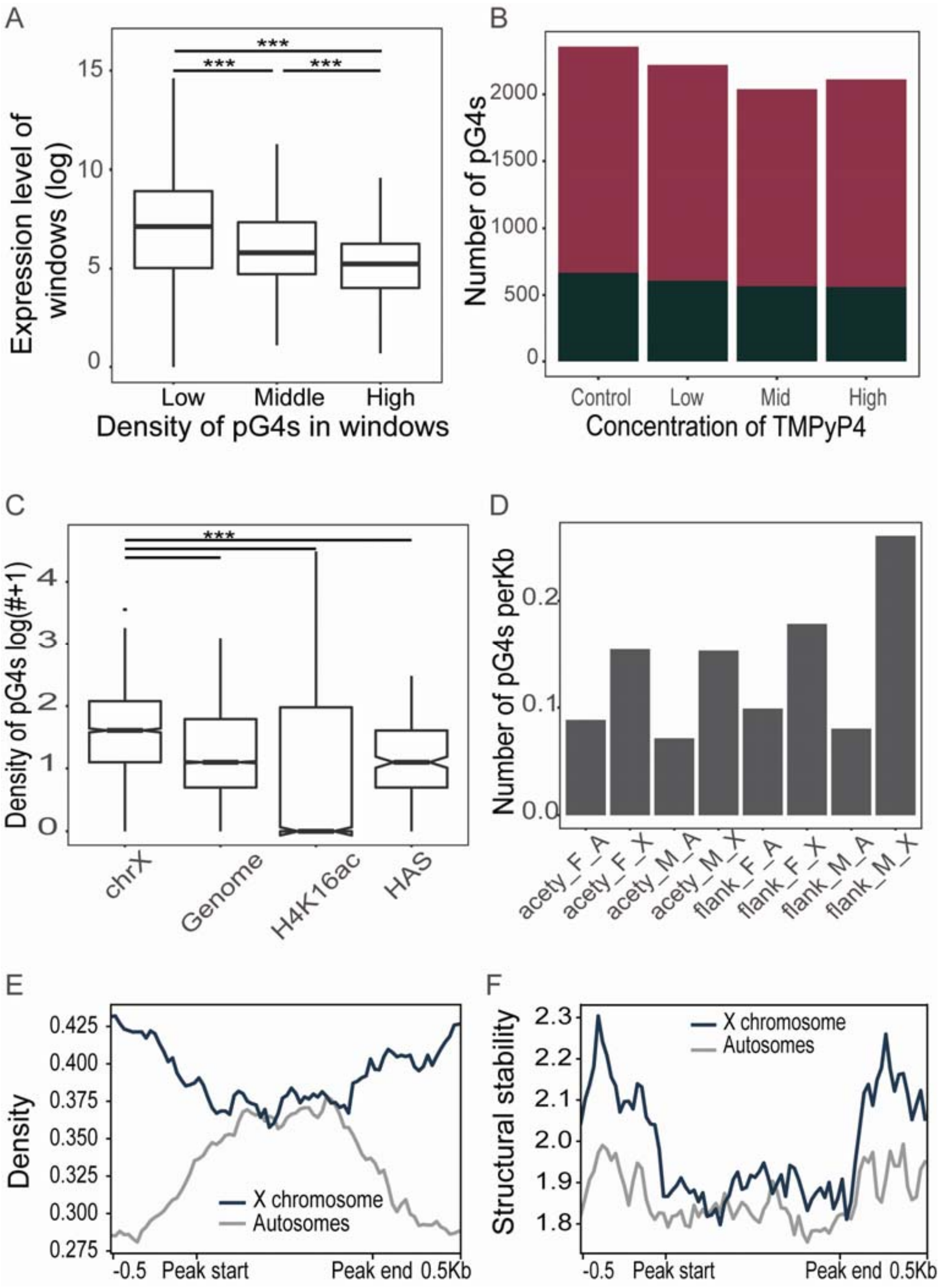

Fig. 3 The low density and structural stability of G4s on the H4K16ac and HAS regions. (A) The correlation of a window between pG4 density and expression. A sliding window strategy was used to generate all genomic windows (1Kb) and divided into three groups based on thier pG4 density. We found the more pG4s density in a window, the less expression level the window showed. (B) The expression level of pG4s under different concentration of TMPyP4, DMSO was used as control. The pG4s show less expression level when were treated with TMPyP4 with higher concentration. (C) The pG4 density on the X chromosome, the whole genome, the H4K16ac regions and HAS regions. (D) X-linked specific flanking regions of H4K16ac in male showed

the highest density of pG4s than other regions, acety means H4K16ac region, flank means flanking region of H4K16ac. (E and F) Density and structural stability profile of G4s in H4K16ac (scale to 1 kb) and flanking regions (upstream and downstream 500bp) on the male X chromosome and autosomes, respectively. Region from peak start to end indicated H4K16ac region.

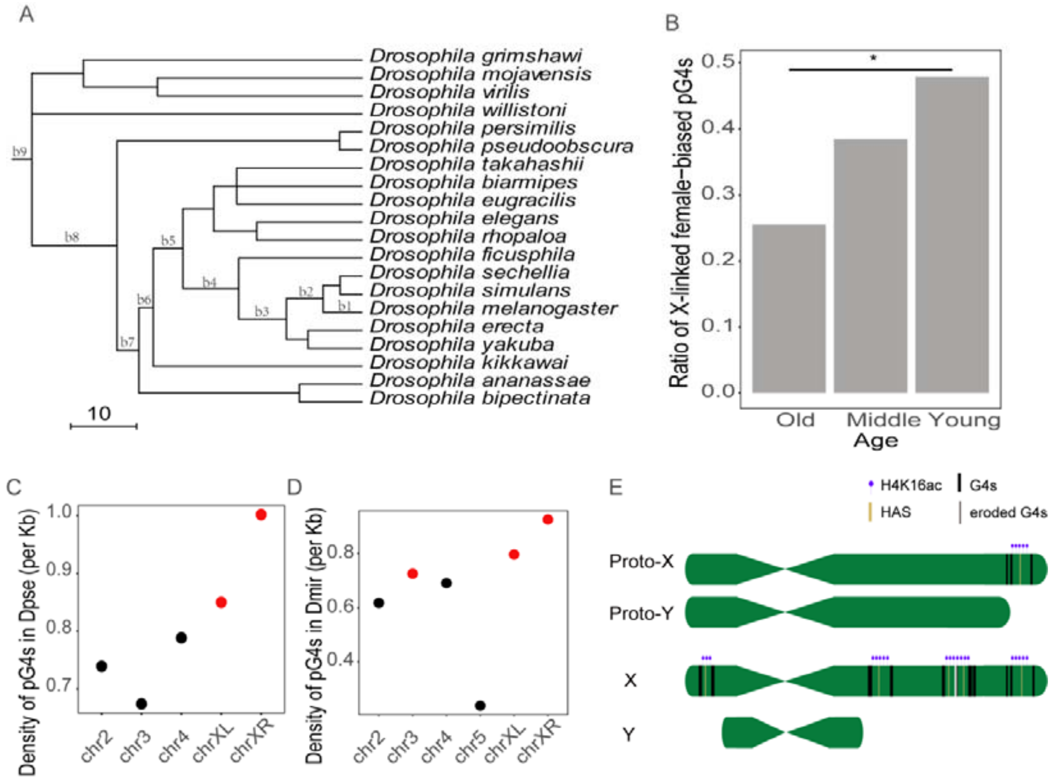

Fig. 4 Gradually accumulated G4s on the X chromosome functions as insulators in dosage compensation. (A) Phylogenetic tree of 20 drosophila species. (B) X-linked G4s with female-biased expression were more likely to be younger. (C and D) Chromosomal density of pG4s in Dpse and Dmir. (E) Model of G4 evolution in dosage compensation. In the early stage of sex chromosome evolution, the Y chromosome still maintained abundant gene homologous to the X chromosome. Only the genes with single dose in male should be compensated, while other two-dose genes were not to be. The gradually formed G-quadruplexes enriched on the flanking regions around MSL binding sites and H4K16 acetylated regions, acting as insulators to precisely separate from the regions that should not be up-regulated. With the degradation of the Y chromosome, the X chromosome acquired more and more high-affinity sites (HAS) for large-scale dosage compensation, and the G4s located between two or many HAS may be eroded extend compensation regions.