GA-repeats on mammalian X chromosomes support Ohno's hypothesis of dosage compensation by transcriptional upregulation

Edridge D'Souza[1], Elizaveta Hosage[1], Kathryn Weinand[2], Steve Gisselbrecht[2]., Vicky Markstein[3], Peter Markstein[3], Martha L. Bulyk[2], Michele Markstein[1*]

[1]Biology Department, University of Massachusetts Amherst, 611 North Pleasant Street, Amherst MA 01003

[2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115

[3]in silico Labs, 160 Redland Road, Woodside, CA 94062

***Corresponding Author:**

Michele Markstein

Email: mmarkstein@bio.umass.edu

Phone: 617-459-9002

ABSTRACT

Over 50 years ago, Susumo Ohno proposed that dosage compensation in mammals would require upregulation of gene expression on the single active X chromosome, a mechanism which to date is best understood in the fruit fly *Drosophila melanogaster*. Here, we report that the GA-repeat sequences that recruit the conserved MSL dosage compensation complex to the Drosophila X chromosome are also enriched across mammalian X chromosomes, providing genomic support for the Ohno hypothesis. We show that mammalian GA-repeats derive in part from transposable elements, suggesting a mechanism whereby unrelated X chromosomes from dipterans to mammals accumulate binding sites for the MSL dosage compensation complex through convergent evolution, driven by their propensity to accumulate transposable elements.

RESULTS AND DISCUSSION

Dosage compensation in placental mammals has long been known to involve epigenetic silencing of one of the two X chromosomes in females[1], resulting in one silenced X chromosome and one active X chromosome. This mechanism ensures that the active X to autosome ratio is the same between females, which have two X chromosomes, and males, which have only one X chromosome in each somatic cell. In addition, a growing body of evidence suggests that dosage compensation in mammals involves a second mechanism that augments gene expression from the single active X chromosome so that it is equivalent to the output of two active X chromosomes[2–9]. This second mechanism was first proposed in 1967 by Susumu Ohno,[10] who argued that increased expression from the X chromosome would be required to avoid the consequences of aneuploidy arising from the evolutionary degeneration of its homolog into a gene-poor Y chromosome. However, while X-linked sequences that mediate the silencing arm of dosage compensation, such as Xist, have been identified, X-linked sequences that target genes on the active X chromosome for augmented expression have remained elusive.

Augmented expression of X-linked genes occurs by two mechanisms in mammalian cells: increased transcription and increased stabilization of RNA transcripts[7,9,11]. It is generally thought that these mechanisms evolved on a gene-by-gene basis on the X chromosome to compensate for the loss of homologous genes on the Y chromosome[7,12]. Consistent with this model, gene-specific microRNA target sequences correlate with the "dosage-sensitivity" of X-linked genes[13]. However, to date no DNA sequences have been reported that support either a gene-by-gene or X-chromosome wide mechanism resulting in augmented transcription of dosage-compensated genes in mammals.

Mechanistically, transcriptional upregulation of X-linked genes has been linked to the activity of the acetyltransferase, MOF/KAT8, which acetylates histone 4 at lysine 16 (H4K16-ac), resulting in open chromatin and increased transcription[14,15]. ChIP-seq experiments in mammalian cells show that MOF/KAT8, H4K16-ac, and RNA pol II are enriched about two-fold at upregulated X-linked genes relative to autosomal genes[7,9]. Moreover, RNAi knockdown of MOF/KAT8 in mammalian cells reduces the two-fold enrichment of RNA Pol II at several X-linked genes and their levels of transcription[9]. The question of how X-linked genes become targeted for transcriptional upregulation is therefore tied to how MOF/KAT8 becomes enriched at X-linked genes.

Since MOF/KAT8 plays a similar role in dosage compensation in the fruit fly *Drosophila melanogaster*[16,17]*,* we reasoned that its mechanism of enrichment might lend insight into how X-linked genes are targeted for upregulation in mammals. In both flies and mammalian cells, MOF/KAT8 is recruited to the X chromosome as part of the MSL dosage compensation complex, which contains the conserved proteins, MSL1, MSL2, and MSL3[14–17]. In flies, these proteins have been shown to direct the MSL complex to the X chromosome in a two-step process: MSL1 and MSL2 are required for recruitment of the complex to about 300 “chromatin entry sites” (CES), also called high affinity sites (HAS) along the X chromosome, and MSL3, a chromodomain protein, is required for spreading of the complex from the CES/HAS sites to neighboring genes[18–20]. In addition, two zinc finger proteins, CLAMP and GAF, recruit the MSL complex to DNA[21,22]. CLAMP acts locally at CES/HAS sites and at a distance with GAF to recruit the MSL complex and to shape the overall architecture of the X chromosome[22–24]. While it is unclear if there is a CLAMP homolog in mammals, molecular modeling has identified c-krox8/Th-POK as a mammalian GAF homolog[25].

Analysis of the 300 recruitment CES/HAS sites in Drosophila, as well as in vitro binding assays with MSL2, CLAMP, and GAF, point to the importance of GA-dinucleotides in recruiting the MSL complex to the X chromosome. For example, analysis of the CES/HAS sites identified a 21-bp “MSL Recognition Element” (MRE) containing 8-bp GA-repeat core that is necessary for MSL complex recruitment[18]. In vitro binding studies show that MSL2 binds an MRE-like sequence, called “PionX” that while different from the 5’ and 3’ ends of the MRE, retains the 8-bp GA-repeat core[26,27]. Additionally, in vitro binding assays with CLAMP show preferential binding to longer GA-repeats between 10 and 30 bp[23].

To determine if the density of GA-repeats is likewise enriched on mammalian X chromosomes and at mammalian X-linked genes, we developed “GenomeHash”, an algorithm to count user-defined motifs throughout specified genomes (see Methods). We validated our algorithm against the Drosophila genome, confirming previously published findings that the densities of GA-repeats are enriched ≥2-fold on the X chromosome relative to autosomes (Supplemental Table I). For example, GA-repeat lengths from 8 bp to 28 bp, experimentally validated in MSL and CLAMP binding assays in vitro and in vivo[18,23,27], occur on the X chromosome at densities ≥1.0/Mb, with an average X:A density enrichment of 2.5-fold (Figure 1a).

We found that the density of GA-repeats on the human X chromosome is likewise enriched relative to autosomes. Most prominently, GA-repeats of lengths 18–38 bp occur on the X chromosome at densities ≥ 1.0/Mb where they are enriched 1.5-fold relative to autosomes (Figure 1b, Supplemental Table 1). Although the X:A enrichment of GA-repeat densities is more modest than observed in Drosophila, it is statistically significant based on both Poisson tests and empirical bootstrapping methods (see Materials and Methods) with p-values ranging from 1.84e-57 for 18-mers to 1.82e-07 for 38-mers (Supplemental Table 1). Moreover, we found that dosage compensated genes[28] are more likely than autosomal genes to contain intronic GA-repeats (Supplemental Tables 2 and 3). For example, GA-repeats of length 30 bp occur in 12% of dosage compensated genes but only 5.3% of autosomal genes, resulting in a 2.2-fold enrichment of dosage compensated genes. We found similar results with clusters of GA-rich consensus motifs matching the Drosophila MSL recognition element (MRE), the

Drosophila CLAMP protein, and the mammalian GAF homolog (Supplemental Tables 2). Collectively, these findings show that the density and distribution of GA-repeats on the human X chromosome, as in Drosophila, are compatible with mediating chromosome-wide and gene-by-gene mechanisms of dosage compensation.

As further confirmation of the possibility that GA-repeats mediate dosage compensation in mammals, we found that GA-repeats are enriched on the X-chromosome not only in the human genome but also the chimpanzee, gorilla, dog, cat, cow, horse, mouse, rat, and opossum genomes (Supplemental Table 1). Focusing on 20-mer GA repeats, two patterns of enrichment are apparent (Figure 1c). One pattern follows the ~1.5-fold X:A density enrichment of GA-repeats that is observed in humans, and is evident in dogs, primates, horses, and cows. The second pattern is characterized by high densities of GA-repeats genome-wide. These higher densities of GA-repeats are about an order of magnitude greater than found on the Drosophila X chromosome and are evident on the X chromosome and autosomes of opossum, mouse, rat, cat and dog. In general, highly dense genome-wide occurrences of GA-repeats precludes their enrichment on the X chromosome relative to autosomes, as observed for the near-ubiquitous occurrences of shorter GA-repeats in flies and humans. However, as observed in dogs, it is possible for a genome to exhibit both patterns of GA-repeat enrichment.

Importantly, the densities of long GA-repeats across mammalian X chromosomes are statistically significant by Poisson tests and empirical bootstrapping methods (Supplemental Tables 2 and 3) and on par with other biologically active motifs that shape whole chromosomes. In fact, the density of 20 bp GA-repeats on all 10 mammalian X chromosomes is more dense than the well-characterized genome-wide insulator protein, CTCF, which occurs about once every million base pairs across human chromosomes[29]. Additionally, 8 of the 10 mammalian genomes we examined exhibit X chromosome GA-repeat densities on par with or greater than the density of GA-repeats associated with dosage compensation in Drosophila[23].

Our finding that GA-repeats are abundant on X chromosomes across mammals supports the hypothesis that the MSL dosage compensation complex, known to interact with GA-repeats on the X-chromosomes in dipterans to augment gene expression, likewise engages with GA-repeats on mammalian X chromosomes to augment gene expression. However, unlike the core proteins of the MSL complex, which are derived from the last common ancestor between mammals and dipterans, GA-repeats cannot have arisen from a common ancestor, as the X chromosomes in dipterans and mammals have completely different evolutionary histories. In fact, even within dipterans the X chromosomes have different evolutionary histories[30], and have been shown in one case to acquire 8-bp GA-repeats through invasion and domestication of a transposable element (TE)[31].

We likewise identified several examples of TEs associated with GA-repeats in the human genome. For example, LINE subfamilies L1 and L2, and the SINE subfamilies AluJ, AluS, AluY, have contributed loci with tandem duplications[32] containing GA-repeats ranging from 8 bp to 28 bp (Supplemental Table 4). These findings suggest that further investigation of LINES, some of which exhibit a 2-fold X:A enrichment[33] and SINES, present in high copy numbers in the genomes of opossum, mouse, rat, and dog, may explain the two patterns of GA-repeat enrichment we observed in mammalian genomes (Figure 1c). Additionally, other TE families may have also contributed to GA-repeat enrichment on the X chromosome. For example, we discovered that the mammalian gypsy retrotransposon is a good candidate: it contains clusters of GA-rich consensus sequences[35] ranging from 6-12 bp in their Long Terminal Repeats and is three-fold enriched on the human X chromosome relative to autosomes (p=6.14e-23) (Figure 2). These findings suggest systematic exploration of TEs is likely to shed light on the evolution of GA-repeats in mammalian genomes.

Collectively, our results show that just as the MSL dosage compensation complex is conserved from flies to mammals, its corresponding GA-rich binding core is enriched on the X chromosomes of flies and mammals by convergent evolution. Our finding that GA-repeats in the human genome are derived in part from TEs suggests that any chromosome with a propensity to accumulate TEs and repetitive elements, such as ancient and nascent X chromosomes[31,33,36], is poised to be targeted by transcriptional machinery such as the MSL complex which is recruited by repetitive elements. In fact, we would predict that the MSL complex, by virtue of its recruitment to chromosomes by GA-repeats, is poised to target future X chromosomes, as our results show that accumulation of GA-repeats is a common feature of X chromosomes.

Methods

Software, Statistical Methods, and Databases employed are provided in the Supplemental Materials and Methods

Acknowledgements

We are indebted to Erica Larschan for insightful conversations and Jake Mayfield, Laura Quilter, Rachel Brody and Kristopher Kolbert for critically reviewing the manuscript.

Author contributions

M.M. and E.D. conceived the project; M.M., E.H., E.D., K.W., S.G., M.B. designed the experiments; E.D., E.H., K.W., S.G. conducted the experiments; V.M. and P.M. wrote GenomeHash; E.D. generated the figures; E.D. and K.W. wrote materials and methods; M.M. wrote the manuscript; M.B. and M.M. supervised the project.

References

1. Lyon, M. F. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* **190**, 372-373 (1961).

2. Nguyen, D. K. & Disteche, C. M. Dosage compensation of the active X chromosome in mammals. *Nat Genet* **38**, 47-53 (2006).

3. Gupta, V. et al. Global analysis of X-chromosome dosage compensation. *J Biol* **5**, 3 (2006).

4. Lin, H. et al. Dosage compensation in the mouse balances up-regulation and silencing of X-linked genes. *PLoS Biol* **5**, e326 (2007).

5. Deng, X. et al. Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster. *Nat Genet* **43**, 1179-1185 (2011).

6. Kharchenko, P. V., Xi, R. & Park, P. J. Evidence for dosage compensation between the X chromosome and autosomes in mammals. *Nat Genet* **43**, 1167-9; author reply 1171 (2011).

7. Yildirim, E., Sadreyev, R. I., Pinter, S. F. & Lee, J. T. X-chromosome hyperactivation in mammals via nonlinear relationships between chromatin states and transcription. *Nat Struct Mol Biol* **19**, 56-61 (2011).

8. Pessia, E., Makino, T., Bailly-Bechet, M., McLysaght, A. & Marais, G. A. Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. *Proc Natl Acad Sci U S A* **109**, 5346-5351 (2012).

9. Deng, X. et al. Mammalian X upregulation is associated with enhanced transcription initiation, RNA half-life, and MOF-mediated H4K16 acetylation. *Dev Cell* **25**, 55-68 (2013).

10. Ohno, S. *Sex chromosomes and sex-linked genes.* (Springer-Verlag, Berlin, New York [etc.], 1967).

11. Lucchesi, J. C. Transcriptional modulation of entire chromosomes: dosage compensation. *J Genet* **97**, 357-364 (2018).

12. Deng, X., Berletch, J. B., Nguyen, D. K. & Disteche, C. M. X chromosome regulation: diverse patterns in development, tissues and disease. *Nat Rev Genet* **15**, 367-378 (2014).

13. Naqvi, S., Bellott, D. W., Lin, K. S. & Page, D. C. Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution. *Genome Res* **28**, 474-483 (2018).

14. Smith, E. R. et al. A human protein complex homologous to the Drosophila MSL complex is responsible for the majority of histone H4 acetylation at lysine 16. *Molecular and cellular biology* **25**, 9175-9188 (2005).

15. Keller, C. I. & Akhtar, A. The MSL complex: juggling RNA–protein interactions for dosage compensation and beyond. *Current opinion in genetics & development* **31**, 1-11 (2015).

16. Kind, J. et al. Genome-wide analysis reveals MOF as a key regulator of dosage compensation and gene expression in Drosophila. *Cell* **133**, 813-828 (2008).

17. Gelbart, M. E., Larschan, E., Peng, S., Park, P. J. & Kuroda, M. I. Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. *Nature structural & molecular biology* **16**, 825 (2009).

18. Alekseyenko, A. A. et al. A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome. *Cell* **134**, 599-609 (2008).

19. Sural, T. H. et al. The MSL3 chromodomain directs a key targeting step for dosage compensation of the Drosophila melanogaster X chromosome. *Nature structural & molecular biology* **15**, 1318 (2008).

20. Straub, T., Grimaud, C., Gilfillan, G. D., Mitterweger, A. & Becker, P. B. The chromosomal high-affinity binding sites for the Drosophila dosage compensation complex. *PLoS genetics* **4**, e1000302 (2008).

21. Larschan, E. et al. Identification of chromatin-associated regulators of MSL complex targeting in Drosophila dosage compensation. *PLoS genetics* **8**, e1002830 (2012).

22. Soruco, M. M. L. et al. The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. *Genes & development* **27**, 1551-1556 (2013).

23. Kuzu, G. et al. Expansion of GA dinucleotide repeats increases the density of CLAMP binding sites on the X-chromosome to promote Drosophila dosage compensation. *PLoS genetics* **12**, e1006120 (2016).

24. Kaye, E. G. et al. Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the Drosophila Dosage Compensation Complex to the Male X Chromosome. *Cell Rep* **22**, 3227-3239 (2018).

25. Matharu, N. K., Hussain, T., Sankaranarayanan, R. & Mishra, R. K. Vertebrate homologue of Drosophila GAGA factor. *J Mol Biol* **400**, 434-447 (2010).

26. Zheng, S. et al. Structural basis of X chromosome DNA recognition by the MSL2 CXC domain during Drosophila dosage compensation. *Genes Dev* **28**, 2652-2662 (2014).

27. Villa, R., Schauer, T., Smialowski, P., Straub, T. & Becker, P. B. PionX sites mark the X chromosome for dosage compensation. *Nature* **537**, 244 (2016).

28. Tukiainen, T. et al. Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244-248 (2017).

29. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211 (2009).

30. Vicoso, B. & Bachtrog, D. Numerous transitions of sex chromosomes in Diptera. *PLoS Biol* **13**, e1002078 (2015).

31. Ellison, C. E. & Bachtrog, D. Dosage compensation via transposable element mediated rewiring of a regulatory network. *Science* **342**, 846-850 (2013).

32. Ahmed, M. & Liang, P. Transposable elements are a significant contributor to tandem repeats in the human genome. *Comp Funct Genomics* **2012**, 947089 (2012).

33. Ross, M. T. et al. The DNA sequence of the human X chromosome. *Nature* **434**, 325-337 (2005).

34. Schmidt, D. et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335-348 (2012).

35. Hubley, R. et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**, D81-9 (2016).

36. Joshi, S. S. & Meller, V. H. Satellite Repeats Identify X Chromatin for Dosage Compensation in Drosophila melanogaster Males. *Curr Biol* **27**, 1393-1402.e2 (2017).

Figures and Figure Legends

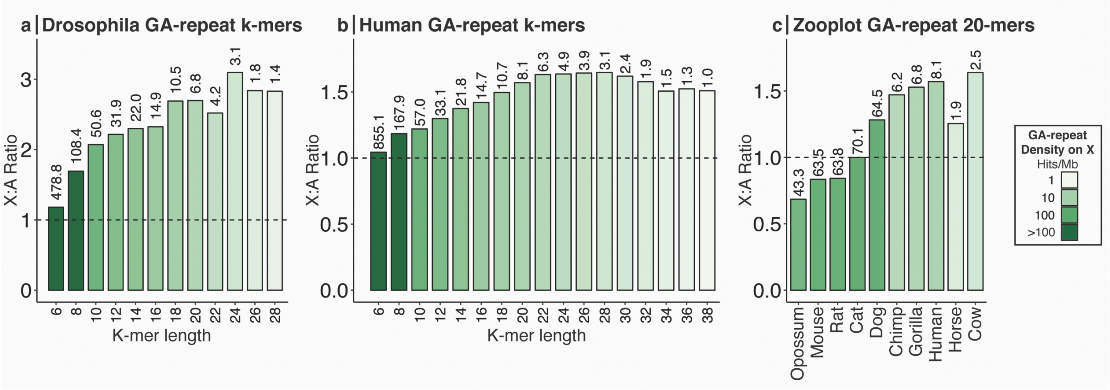

Fig.1| Density of GA-repeats on the X chromosomes of flies and mammals The density of GA-repeats per megabase (Mb) was computed on each chromosome by dividing the number of GA-repeat matches (hits) by the length of each chromosome. The Y-axis shows the ratio of GA-repeat densities on the X chromosome vs. autosomes for GA-repeats of specific lengths. Numbers above each bar represent the density of matches for each GA-repeat k-mer on the X chromosome, which as shown by the color-key, cluster into 3 orders of magnitude. **a**. GA-repeat k-mers in the *Drosophila melanogaster* genome with an X chromosome density ≥ 1 hit/Mb show an average 2.5-fold X:A density enrichment. B. GA-repeat k-mers in the human genome with an X chromosome density ≥ 1 hit/Mb show an average 1.5-fold X:A density enrichment. **c.** The density of GA-repeats of length 20 bp across different mammalian genomes. Genomes with low X:A enrichment ratios tend to have high baseline densities of GA-repeats.

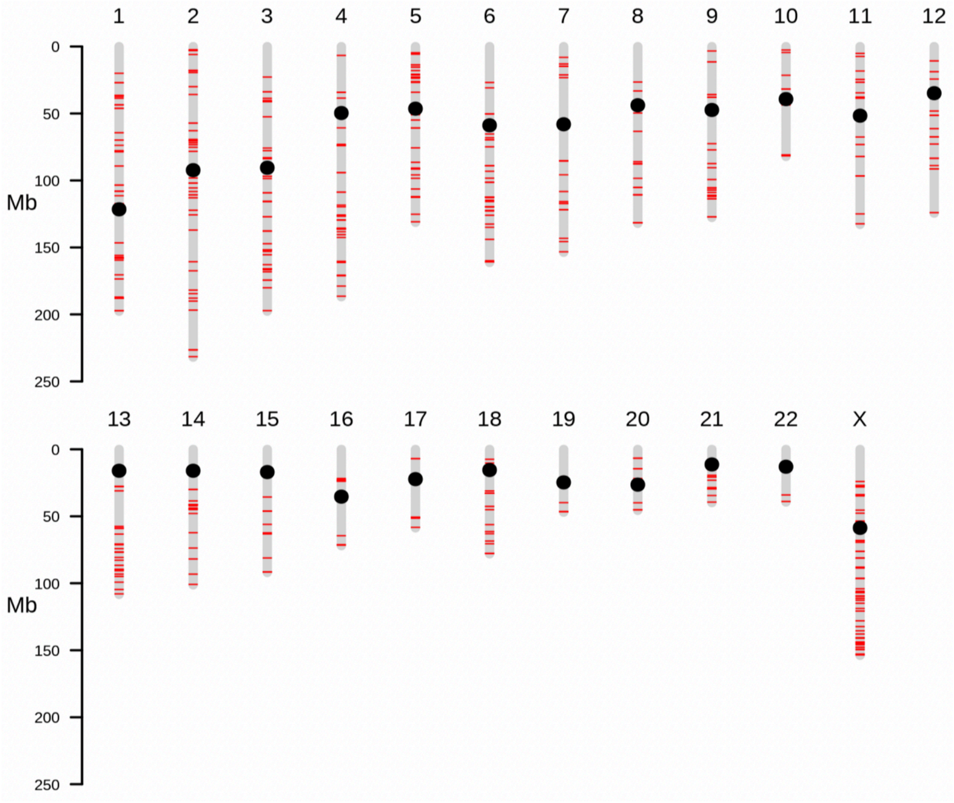

Fig.2| Distribution of the mammalian Gypsy retrotransposon in the human genome. Ideogram of genomic locations matching the Dfam database mammalian gypsy consensus model. The density of matches on the X chromosome is 3-fold higher than the mean autosomal density (p=6.14e-23, 1-tailed upper Poisson test).