Non-allelic gene conversion enables rapid evolutionary change at multiple regulatory sites encoded by transposable elements

Christopher E. Ellison & Doris Bachtrog

Department of Integrative Biology, University of California Berkeley, Berkeley, CA 94720, USA

Transposable elements (TEs) allow rewiring of regulatory networks, and the recent amplification of the ISX-element dispersed 77 functional but suboptimal binding-sites for the dosage-compensation-complex to a newly-formed X-chromosome in Drosophila. Here we identify two linked refining-mutations within ISX that interact epistatically to increase binding affinity to the dosage-compensation-complex. Selection has increased the frequency of this derived haplotype in the population, which is fixed at 30% of ISX-insertions and polymorphic among another 41%. Sharing of this haplotype indicates that high levels of gene-conversion among ISX-elements allow them to "crowd-source" refining-mutations, and a refining-mutation that occurs at any single ISX-element can spread in two dimensions: horizontally across insertion sites by non-allelic gene-conversion, and vertically through the population by natural selection. These findings describe a novel route by which fully functional regulatory elements can arise rapidly from TEs and implicate non-allelic gene-conversion as having an important role in accelerating the evolutionary fine-tuning of regulatory networks.

Introduction

A substantial portion of animal genomes is composed of repetitive sequences, including gene duplicates, satellite DNA, and transposable elements. Gene conversion is a major force shaping the evolution of repetitive regions, and interlocus or non-allelic gene conversion between sequence duplicates has been studied extensively for its role in concerted evolution (Chen et al., 2007; Ohta, 2010). Non-allelic gene conversion also affects selection operating in gene families. Compared to single-copy genes, a family of gene duplicates presents a larger mutational target, and a mutation arising in any gene copy can be spread among copies by non-allelic gene conversion, thereby increasing the efficiency of both positive and purifying selection (Mano and Innan, 2008). Non-allelic gene conversion homogenizes the arrays of ribosomal DNA gene

copies present in the genomes of most organisms (Eickbush and Eickbush, 2007), has generated allelic diversity within the human leukocyte antigen gene family (Zangenberg et al., 1995), and has allowed palindromic genes on the human Y chromosome to escape degeneration (Rozen et al., 2003).

Transposable elements give rise to families of duplicate sequences. A propensity for some TEs to carry regulatory motifs and to insert adjacent to coding sequence gives them the potential for being potent modulators of gene regulatory networks (Cowley and Oakey, 2013; Feschotte, 2008). The regulatory elements provided by these TEs, however, may be suboptimal in function, and subject to subsequent fine-tuning (Polavarapu et al., 2008). Unlike regulatory elements where short binding motifs (ten basepairs on average for transcription factors; Stewart et al., 2013) evolve *de novo* via point mutation or microsatellite expansion, binding sites that evolve from TEs are initially almost identical in sequence and are nested within a larger repeat unit (hundreds or thousands of basepairs in size), and may thus be subject to non-allelic gene conversion. Re-wiring of the dosage compensation network in *Drosophila miranda* was driven by TE-mediated amplification of a functional but suboptimal binding motif (Ellison and Bachtrog, 2013). Here we show that non-allelic gene conversion is catalyzing the rapid fine-tuning of these suboptimal motifs by allowing sequence variants that optimize binding affinity to spread among elements.

Dosage compensation in Drosophila is mediated by a male-specific ribonucleoprotein complex (the male-specific lethal or MSL complex) that binds to a GA-rich sequence motif (the MSL recognition motif) at a number of chromatin entry sites on the X chromosome (Alekseyenko et al., 2008; Straub et al., 2008). We previously studied the acquisition of novel chromatin entry sites on newly formed X chromosomes in *Drosophila miranda*, a species where two independent sex chromosome/autosome fusions resulted in a karyotype composed of three X chromosome arms, each of a different age (Alekseyenko et al., 2013; Zhou et al., 2013). XL is homologous to the X chromosome of *D. melanogaster* and has been a sex chromosome for at least 60 million years (Richards et al., 2005); chromosome XR formed roughly 15 million years ago when an autosome (Muller element D) fused to XL (Carvalho and Clark, 2005), and the neo-X/neo-Y chromosome pair originated around 1.5 million years ago when the Y fused to another autosome

(Muller element C) (Bachtrog and Charlesworth, 2002). Dosage compensation evolved on both XR and the neo-X shortly after their emergence, through acquisition of novel chromatin entry sites and co-option of the MSL regulatory network (Bone and Kuroda, 1996; Marin et al., 1996). Interestingly, we discovered that the acquisition of dosage compensation on both XR and the neo-X chromosome was in part mediated by the independent domestication of helitron transposable elements that contained MSL recognition motifs, which we have termed ISXR and ISX, respectively (Ellison and Bachtrog, 2013).

ISX is highly enriched on the neo-X chromosome of *D. miranda* and is derived from the abundant ISY element. Compared to ISY, ISX contains a 10 basepair deletion that creates a MSL recognition motif, thereby allowing it to act as a chromatin entry site (Ellison and Bachtrog, 2013). Our previous study showed that while amplification of ISX about 1 million years ago provided dozens of functional chromatin entry sites on the neo-X chromosome of *D. miranda*, the motif dispersed by ISX is distinct from the canonical motif that is enriched within chromatin entry sites on XL and XR, and shows significantly lower affinity to the MSL complex compared to motifs on XL and XR (Ellison and Bachtrog, 2013). For these reasons, we postulated that the ISX binding motif is suboptimal, and predicted that refining mutations should accumulate within each MSL recognition motif until the neo-X chromosome becomes fully dosage compensated (Ellison and Bachtrog, 2013).

Results

Variation at MSL recognition motifs among ISX insertions in *D. miranda* strain MSH22

To identify potential refining mutations that optimize MSL-binding at chromatin entry sites derived from the ISX element, we characterized sequence variation within the MSL recognition motifs and flanking sequence regions for all 77 insertions of the ISX element on the neo-X chromosome in the sequenced reference strain MSH22 (**Figure 1A**). Because we have previously demonstrated that ISX contains a functional MSL recognition motif but the closely related ISY element does not (Ellison and Bachtrog, 2013), we sought to identify sequence variants that were present in multiple ISX elements but rare or absent in ISY elements from the same chromosome.

Using these criteria, we identified a sequence haplotype adjacent to the MSL recognition motif that is common among MSH22 ISX insertions and rare among ISY elements: 57% of ISX elements carry this haplotype versus 0.7% of neo-X ISY insertions, an asymmetry significantly different from that expected by chance (Fisher's Exact Test; $P < 2.2e\text{-}16$). The haplotype consists of two mutations (G → T and A → T), separated by two basepairs, which are in perfect linkage disequilibrium among ISX but not ISY elements (**Figure 1 & Figure 1-figure supplement**). Because ISX is descended from ISY and the TT alleles are rare among ISY elements, they are likely to be derived. We hereafter refer to these mutations as the TT haplotype.

The TT haplotype increases MSL complex binding affinity

To determine if the TT haplotype affects binding affinity of the MSL complex, we used published ChIP-seq data of MSL3 (a component of the MSL complex) from *D. miranda* strain MSH22 (Alekseyenko et al., 2013). We compared *in vivo* MSL complex binding levels for the 44 MSH22 ISX insertions carrying the TT haplotype to the 33 insertions with the ancestral GA haplotype. The insertions with the TT alleles had significantly higher levels of MSL complex binding compared to those with the GA alleles (Wilcoxon test P=0.01; **Figure 2A**).

We previously demonstrated that insertion of an ISX element in the *D. melanogaster* genome results in recruitment of the MSL complex to an ectopic autosomal location (Ellison and Bachtrog, 2013). We used this same system to dissect the relationship between the TT alleles and MSL complex binding affinity. Starting with a cloned ISX element (Ellison and Bachtrog, 2013), we used site-directed mutagenesis to create variants of this element that differ only with respect to the TT haplotype. Each of the four possible haplotypes (GA, GT, TA, and TT) was engineered and inserted onto *D. melanogaster* chromosome 2L at cytosite 38F1 using recombinase mediated cassette exchange (RMCE) (Bateman et al., 2006). We then measured the effect of each of the derived variants by quantifying allele-specific binding levels of the MSL complex in F1 hybrids between the ancestral haplotype (GA) and each of the derived haplotypes (GT, TA, and TT).

Interestingly, each T allele, when assayed separately, has a negative effect on MSL binding levels compared to the ancestral G or A allele (**Figure 2B**). However, when combined, the TT haplotype results in significantly increased levels of MSL complex binding, relative to the

ancestral GA haplotype (Wilcoxon Test $P = 0.0289$; **Figure 2B**). These results suggest that there is sign epistasis between the two alleles and that the high frequency TT haplotype represents a refining/fine-tuning adaptation, since recruitment of MSL complex to the adjacent MSL recognition motif is increased.

Non-allelic gene conversion is spreading the TT haplotype among ISX insertions

It is unlikely that the TT haplotype arose multiple times by parallel mutation, and there are two possibilities that could explain its prevalence among MSH22 ISX insertions. First, this double mutation may have occurred early during the process of ISX amplification, thus giving rise to two lineages of ISX: one that carries the ancestral GA haplotype, and the other with the TT haplotype. The TT-harboring elements in MSH22 would then all be descendants from the latter ISX lineage. Alternatively, this mutation may have occurred only after the GA-containing ISX element was fixed in the population at all 77 neo-X insertion sites, at which point it was spread among independent ISX elements via non-allelic gene conversion.

We can distinguish between these possibilities by examining patterns of sequence polymorphism for each ISX insertion across multiple strains of *D. miranda*. A canonical signature of non-allelic gene conversion is the presence of shared polymorphisms across sequence duplicates (Arguello et al., 2006; Mansai and Innan, 2010). If gene conversion is spreading the TT haplotype among ISX insertions, we expect it to be polymorphic among individuals at several ISX insertion sites, whereas we do not expect the TT haplotype to be polymorphic at individual ISX insertions under the alternative scenario.

To genotype multiple wild-derived individuals at each of the MSH22 ISX insertions, we used paired-end Illumina genomic resequencing data from 23 inbred lines of *D. miranda*, including MSH22. We aligned all reads to the MSH22 reference genome and identified mate-pairs where one mate was anchored in unique sequence flanking an ISX insertion. We then assembled these reads to generate a contig spanning the 5’ flank of the ISX element insertion, which contains the MSL recognition motif, for each inbred line. Using this approach we generated population data for 69 insertions out of the 77 total ISX insertions present in the MSH22 reference genome assembly. Uneven sequence coverage between insertions and individuals meant that not all

insertions could be assembled for each individual. However, the majority of individuals are represented in the majority of datasets: each insertion dataset contained ~20 lines on average (see Dataset S1 in Dryad: Ellison & Bachtrog, 2015). Almost all ISX insertions are fixed among strains (68 of 69) and insertion sites are identical between lines, suggesting that independent parallel insertions are unlikely to be present within our dataset. We performed PCR and Sanger sequencing on a subset of these regions and estimate the base-calling error rate of our Illumina contigs to be ~0.1%.

Consistent with non-allelic gene conversion spreading the TT haplotype, we observe a strong signal of allele-sharing within the sequence region flanking the MSL recognition motif among ISX insertions (**Figure 3**). On average, 68.9% of polymorphisms observed within a given insertion are shared among other insertions (though most polymorphisms are shared only between a few elements). The TT haplotype is especially striking in this regard as it is polymorphic in 41% of insertions (**Figures 3, 4, & Figures 3-figure supplement**). If population subdivision contributes to this excess of allele sharing, we would expect individuals to cluster by allele state at the TT locus, across all polymorphic ISX insertions. Instead, we find that different individuals contribute to the TT polymorphism at each of these ISX insertions (**Figure 4-figure supplement 1**), suggesting that abundant non-allelic gene conversion is the most likely explanation for this observation. Interestingly, the population frequency of the TT haplotype is similar among insertions that are near each other on the chromosome (permutation test P=0.018; **Figure 4**). This is consistent with higher gene conversion rates between more closely linked ISX elements generating correlated population frequencies among adjacent elements (Sasaki et al., 2010).

Selection is driving the spread of the TT haplotype through the population

To test if selection has acted to increase the frequency of the TT haplotype in the population, we examined patterns of polymorphisms at GA- and TT-containing ISX elements. The TT haplotype harbors significantly less linked variation than the ancestral GA haplotype, across insertion sites and individuals (haplotype diversity = 0.53 vs. 0.81; resampling P<0.001; **Figure 5A**). In addition, ISX insertions where TT is fixed have significantly lower nucleotide diversity compared to the insertions where GA is fixed (one-sided Wilcoxon test P=0.035; **Figure**

5B). Finally, the frequency spectrum at the TT haplotype also shows an excess of high frequency derived alleles, compared to the frequency spectrum at the GA haplotype (resampling P=0.027; **Figure 5C**). All of these patterns are expected if natural selection acting on the TT haplotype is driving its spread through the population.

Discussion

Recent work in a variety of eukaryotes suggests that transposable elements may be major drivers of regulatory evolution (Cowley and Oakey, 2013; Feschotte, 2008). Their high transposition rate and ability to supply ready-to use regulatory elements across the genome implies that they may rapidly wire new genes into regulatory networks (Feschotte, 2008). We recently showed that domesticated TEs contribute to rewiring of the dosage compensation network in *D. miranda*, but appear to supply only suboptimal binding sites for the MSL complex (Ellison and Bachtrog, 2013). Here, we identify a derived haplotype with two mutations that interact epistatically to increase binding affinity for the MSL complex. We show that these fine-tuning mutations spread among independent ISX insertions by non-allelic gene conversion, and through the population by natural selection (**Figure 6**). Relative to regulatory elements that evolve in isolation, a family of regulatory motifs dispersed by TEs presents a larger mutational target, and a mutation arising in any element contained within a larger repeat unit (the TE) can spread among copies by non-allelic gene conversion. Consequently, the rate of evolutionary fine-tuning at such regulatory elements can be greatly accelerated by increasing their effective population size (Mano and Innan, 2008). Thus, transposable elements can "crowd-source" beneficial mutations to rapidly fine-tune regulatory networks.

Our transgenic experiments show that each individual T allele actually decreases the binding affinity for the MSL complex relative to the ancestral GA haplotype. Thus, TA or GT haplotypes should be selected against in the population if present on a functional ISX element. Consistent with the deleterious effect of individual T alleles, the TA and GT haplotypes are present on some ISY elements but completely absent from ISX, i.e., we find the two T mutations to be in perfect linkage disequilibrium among ISX elements but not ISY (**Figure 1B**). While most ISY elements carry the ancestral GA haplotype, a small fraction (0.7% of neo-X ISY insertions) instead carry the derived TT haplotype. It is therefore possible that the TT haplotype was introduced onto the

ISX background by non-allelic gene conversion from ISY. Under this scenario, the large family of ISY elements in the *D. miranda* genome could be acting as a reservoir of natural variation, where complex mutations can accumulate in the absence of epistasis. Non-allelic gene conversion could then transfer these haplotypes to related repetitive elements (such as ISX). While many of these haplotypes are likely to be neutral or deleterious, some may be beneficial, as in the case of the TT haplotype. Such a scenario avoids the waiting time for a double mutation, as well as the fitness valley that would have to be traversed if the two mutations were to occur sequentially on the ISX background.

To conclude, our findings suggest that TE-dispersed binding motifs follow an evolutionary trajectory that is fundamentally different from those that arise by other means. The complementary roles of TEs in dispersing regulatory motifs, and gene conversion in spreading subsequent refining mutations, combine to allow for the rapid rewiring and fine-tuning of gene regulatory networks. This process adds a new layer of complexity onto how TEs influence regulatory innovation, as well as a new context in which gene conversion affects genome evolution.

References

Alekseyenko, A.A., Ellison, C.E., Gorchakov, A.A., Zhou, Q., Kaiser, V.B., Toda, N., Walton, Z., Peng, S., Park, P.J., Bachtrog, D., *et al.* (2013). Conservation and de novo acquisition of dosage compensation on newly evolved sex chromosomes in Drosophila. Genes Dev *27*, 853-858.

Alekseyenko, A.A., Peng, S., Larschan, E., Gorchakov, A.A., Lee, O.K., Kharchenko, P., McGrath, S.D., Wang, C.I., Mardis, E.R., Park, P.J., *et al.* (2008). A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome. Cell *134*, 599-609.

Arguello, J.R., Chen, Y., Yang, S., Wang, W., and Long, M. (2006). Origination of an X-linked testes chimeric gene by illegitimate recombination in Drosophila. PLoS Genet *2*, 745-754.

Bachtrog, D., and Charlesworth, B. (2002). Reduced adaptation of a non-recombining neo-Y chromosome. Nature *416*, 323-326.

Bateman, J.R., Lee, A.M., and Wu, C.T. (2006). Site-specific transformation of Drosophila via phiC31 integrase-mediated cassette exchange. Genetics *173*, 769-777.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature (London) *456*, 53-59.

Bone, J.R., and Kuroda, M.I. (1996). Dosage compensation regulatory proteins and the evolution of sex

chromosomes in Drosophila. Genetics *144*, 705-713.

Bradley, R.K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., and Pachter, L. (2009). Fast statistical alignment. PLoS Comput Biol *5*, e1000392.

Carvalho, A.B., and Clark, A.G. (2005). Y chromosome of D. pseudoobscura is not homologous to the ancestral Drosophila Y. Science *307*, 108-110.

Chen, J.-M., Cooper, D.N., Chuzhanova, N., Ferec, C., and Patrinos, G.P. (2007). Gene conversion: mechanisms, evolution and human disease. Nature Reviews Genetics *8*, 762-775.

Cowley, M., and Oakey, R.J. (2013). Transposable elements re-wire and fine-tune the transcriptome. PLoS Genet *9*, e1003234.

Eickbush, T.H., and Eickbush, D.G. (2007). Finely orchestrated movements: Evolution of the ribosomal RNA genes. Genetics *175*, 477-485.

Ellison, C.E., and Bachtrog, D. (2013). Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. Science (Washington D C) *342*, 846-850.

Ellison, C.E. and Bachtrog, D. (2015) Data from: Non-allelic gene conversion enables rapid evolutionary change at multiple regulatory sites encoded by transposable elements. Dryad Digital Repository. doi:10.5061/dryad.dg483

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics *9*, 397-405.

Hall, G.S., and Little, D.P. (2007). Relative quantitation of virus population size in mixed genotype infections using sequencing chromatograms. Journal of Virological Methods *146*, 22-28.

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

Mano, S., and Innan, H. (2008). The evolutionary rate of duplicated genes under concerted evolution. Genetics *180*, 493-505.

Mansai, S.P., and Innan, H. (2010). The Power of the Methods for Detecting Interlocus Gene Conversion. Genetics *184*, 517-U292.

Marin, I., Franke, A., Bashaw, G.J., and Baker, B.S. (1996). The dosage compensation system of Drosophila is co-opted by new evolved X chromosomes. Nature *383*, 160-163.

Ohta, T. (2010). Gene conversion and evolution of gene families: an overview. Genes (Basel) *1*, 349-356.

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics (Oxford) *28*, 1420-1428.

Polavarapu, N., Marino-Ramirez, L., Landsman, D., McDonald, J.F., and Jordan, I.K. (2008). Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. BMC Genomics *9*, Article No.: 226.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., *et al.* (2005). Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res *15*, 1-18.

Rozen, S., Skaletsky, H., Marszalek, J.D., Minx, P.J., Cordum, H.S., Waterston, R.H., Wilson, R.K., and Page, D.C.

(2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature (London) *423*, 873-876.

Sasaki, M., Lange, J., and Keeney, S. (2010). Genome destabilization by homologous recombination in the germ line. Nature Reviews Molecular Cell Biology *11*, 182-195.

Stewart, A.J., Hannenhalli, S., and Plotkin, J.B. (2013). Why Transcription Factor Binding Sites Are Ten Nucleotides Long. Genetics *193*.

Straub, T., Grimaud, C., Gilfillan, G.D., Mitterweger, A., and Becker, P.B. (2008). The Chromosomal High-Affinity Binding Sites for the Drosophila Dosage Compensation Complex. PLoS Genet *4*, Article No.: e1000302.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics (Oxford) *25*, 1189-1191.

Zangenberg, G., Huang, M.-M., Arnheim, N., and Erlich, H. (1995). New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. Nat Genet *10*, 407-414.

Zhou, Q., Ellison, C.E., Kaiser, V.B., Alekseyenko, A.A., Gorchakov, A.A., and Bachtrog, D. (2013). The Epigenome of Evolving Drosophila Neo-Sex Chromosomes: Dosage Compensation and Heterochromatin Formation. PLoS Biology *11*, Article No.: e1001711.

Acknowledgments: This work was funded by NIH grants (R01GM076007 and R01GM093182) to D.B. and a NIH postdoctoral fellowship to C.E.E. All DNA-sequencing reads generated in this study are deposited at the National Center for Biotechnology Information Short Reads Archive (**www.ncbi.nlm.nih.gov/sra**) under the BioProject ID PRJNA270105. We thank Molly Przeworski, Jeffrey Fawcett, Isabel Gordo and Monty Slatkin for comments on the manuscript and Daniel Weissman for helpful discussions.

Competing interests: The authors have no competing financial and non-financial interests.

Materials and Methods

Resequencing of *D. miranda* wild lines

Isofemale lines were established from individuals collected in Northern California and inbred for several generations. DNA was extracted from 1-8 females per line using the Qiagen PureGene kit and fragmented by nebulization. Paired-end Illumina libraries were constructed using standard protocols (Bentley et al., 2008) and sequenced on an Illumina Genome Analyzer II machine.

ISX assembly and variant identification

Resequencing data were mapped to version 2.2 of the *D. miranda* MSH22 reference assembly (GenBank: AJMI00000000.2) using bowtie2 (Langmead and Salzberg, 2012). ISX locations were identified in (Ellison and Bachtrog, 2013). Paired-end read alignments were evaluated within 2 kilobase windows flanking each ISX insertion and reads with mapping quality of 20 or greater were extracted along with their mate. The extracted mate pairs were then assembled using IDBA-UD, for each line separately (Peng et al., 2012). Contigs were aligned using FSA (Bradley et al., 2009) and visualized with Jalview (Waterhouse et al., 2009). A custom Perl script (available at https://github.com/chris-ellison/MSAvariants) was used to identify sequence variants within the alignments. We also PCR amplified eight of the ISX insertions where the TT haplotype was polymorphic. We confirmed that this polymorphism was present at each of these insertions and estimated the base-calling accuracy of the assemblies by sequencing the PCR products using Sanger technology.

Transgenesis

We used the QuikChange Lightning site-directed mutagenesis kit from Agilent Technologies and the ISX element cloned in (Ellison and Bachtrog, 2013) to engineer four ISX variants that differed only with respect to the TT haplotype: ISX-GA, ISX-GT, ISX-TA, and ISX-TT. Each construct was injected by BestGene Inc. (Chino Hills) into *D. melanogaster* embryos carrying a RMCE landing site at cytosite 38F1 on chromosome 2L (Bloomington Drosophila Stock Center strain #27388). Transformants were verified by PCR and Sanger sequencing.

Quantification of allele-specific binding levels of the MSL complex

Male third instar larvae (~250 mg) were collected from F1 hybrids between ISX-GA and each of the other three engineered lines: ISX-GT, ISX-TA, and ISX-TT. Chromatin immunoprecipitation was performed for four biological replicates of each of these lines using the MSL2 d-300 primary antibody from Santa Cruz Biotechnology Inc. and the protocol described in (Alekseyenko et al., 2013). Primers flanking the ISX MRE region were used to generate heterozygous amplicons from the MSL2 IP and input control. Sanger chromatograms were used in conjunction with polySNP software (Hall and Little, 2007) to calculate relative abundance of ISX alleles within the IP and input control amplicons. Abundance of the ‘T’ alleles in the IP

amplicons relative to the ancestral G/A alleles was calculated and normalized by the same values from the input control.

Permutation and resampling tests

To determine if the TT frequency among neighboring ISX elements was correlated, we clustered elements within 100kb of each other and calculated the standard deviation in TT allele frequency within clusters. We then compared these values to 1000 permutations where TT allele frequency was randomly shuffled between ISX locations.

The haplotype diversity and allele frequency spectrum resampling tests were performed by drawing, without replacement, two groups of size 617 and 674, respectively, from the pool of 1,291 ISX sequences. The intergroup difference in haplotype diversity, as well as the number of derived variants with frequency of 0.75 or greater, was calculated for each of 1000 replicates and compared to the difference between the TT and GA groups.

Figure Legends

Figure 1. TE-derived MSL recognition element (MRE) motifs from the neo-X chromosome of *Drosophila miranda*

(A) The MSL recognition motif (MRE) plus 20 basepairs of flanking sequence were extracted from all 77 ISX transposable elements located on the neo-X chromosome in the MSH22 reference genome assembly. The multiple sequence alignment of these 77 sequence regions (arranged from top-to-bottom in the order in which they are found on the chromosome) shows that there is sequence variation among elements both within and adjacent to the 21 basepair MRE motif. Each variant has been classified as ancestral or derived based on its frequency in the ISX progenitor element, ISY. The derived allele frequency for each variant in this region is shown for ISX as well as 139 ISY elements from the neo-X chromosome (see **Figure 1-figure supplement** for ISY alignment). Red arrows point to the derived TT haplotype that is common among ISX elements but rare in ISY. (B) Barplot showing the frequencies of all haplotypes at the GA/TT locus, for ISY and ISX elements separately. Two haplotypes are present within ISX elements (GA and TT) and the two alleles within each haplotype are in perfect linkage disequilibrium. In contrast, the majority of ISY elements harbor the GA haplotype, but these two alleles are not in perfect linkage disequilibrium among ISY elements. Rather, five additional allelic combinations are present at low frequencies in this location among ISY, but not ISX elements.

Figure 1 - Supplement. Alignment of ISY elements from the *D. miranda* MSH22 genome assembly

139 ISY elements from the MSH22 neo-X chromosome were identified and 200 basepairs from their 5' flanks were aligned. The black arrows point to the sites where the derived 'T' alleles are common among ISX elements. In contrast, only a single ISY element from the neo-X chromosome harbors the TT haplotype.

Figure 2. The TT haplotype increases MSL binding affinity

(A) MSL3 ChIP-seq data from *D. miranda* strain MSH22 shows that the ISX insertions carrying the TT haplotype recruit significantly higher levels of MSL complex compared to those with the GA haplotype (Wilcoxon test P=0.01). (B) Engineered ISX elements that differ only with respect to the TT haplotype bind different levels of MSL complex. There is an epistatic interaction between the two 'T' alleles such that separately, they decrease MSL complex binding relative to the ancestral allele, but together in the TT haplotype, they increase MSL complex binding (Wilcoxon Test P=0.028 for both comparisons [GT *vs* TT and TA *vs* TT]). The rectangles and error bars show the average and standard deviation of values from four biological replicates for each condition.

Figure 3. ISX variation among wild lines of *D. miranda*

For each ISX insertion identified within the *D. miranda* MSH22 reference genome assembly (alignment shown at left, see also Figure 1), we characterized sequence variation across *D. miranda* individuals. The TT haplotype (magenta lines) was fixed across individuals at 30% of insertions (see example alignment, top right), polymorphic at 41% of insertions (example shown middle right), and absent at 29% of insertions (bottom right). Allele sharing

between insertions occurs at sites other than the TT haplotype, but these sites tend to be shared across fewer insertions (see heatmap, bottom right). **Figure 3-figure supplement** shows the population alignment across all ISX insertions on the neo-X.

Figure 3 - Supplement. Shared polymorphism across sixty-nine ISX insertions

The 5' 200 basepairs of the ISX element was assembled for an average of 20 individuals, for each of 69 ISX insertions. Each stripe corresponds to the population data for a given insertion and nucleotides are colored as in Figure 1. Solid lines point to columns of the alignment containing polymorphisms that are shared between multiple ISX insertions. For these columns, the heatmap is shaded to reflect the degree of allele-sharing, which ranges from 3% of insertions to 41% of insertions. The 'T' letters under the heatmap mark the location of the TT haplotype shown in Figure 1.

Figure 4. Population frequency of TT haplotype across ISX insertions

The location of all ISX elements on the *D. miranda* neo-X chromosome, as inferred from the the MSH22 reference genome assembly, is shown by vertical green bars. The derived TT haplotype (frequency shown in red), is polymorphic at 27 of 69 ISX insertions, a pattern consistent with non-allelic gene conversion.

Figure 4 – Supplement. ISX genotype across insertions and individuals

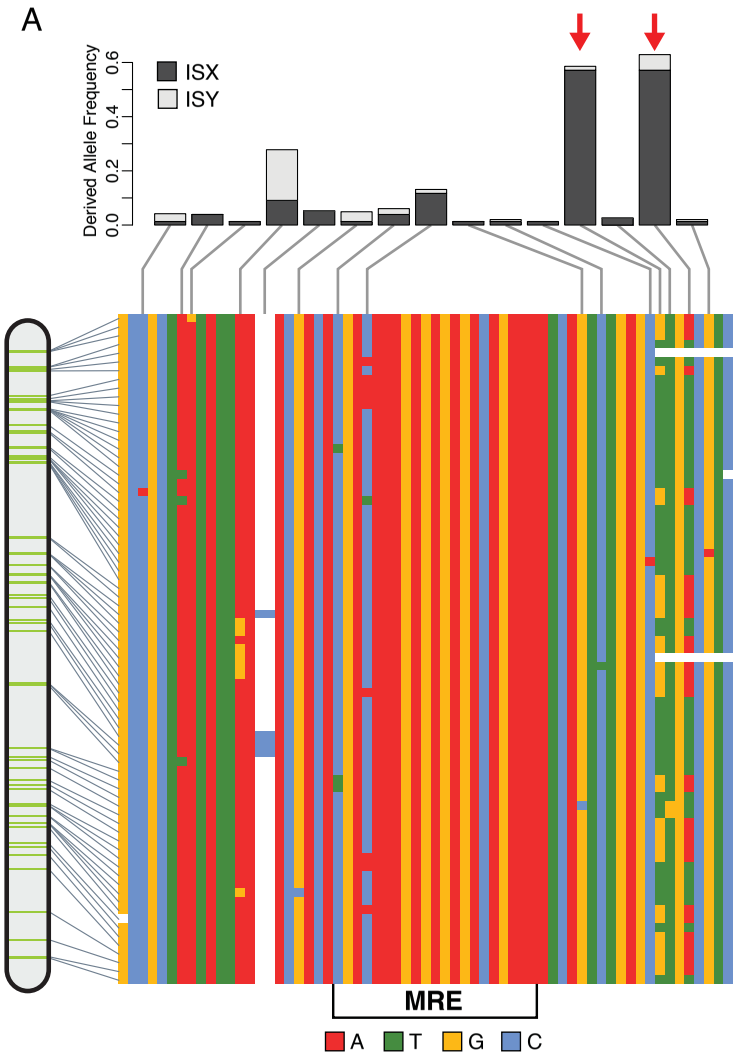
Heatmap showing each of the 27 ISX insertions (rows) where the TT haplotype is polymorphic among individuals. Columns show the genotype of each individual, for each of these insertions. Each individual has a mixture of TT and GA ISX insertions, suggesting that TT polymorphism among lines is not due to population subdivision.

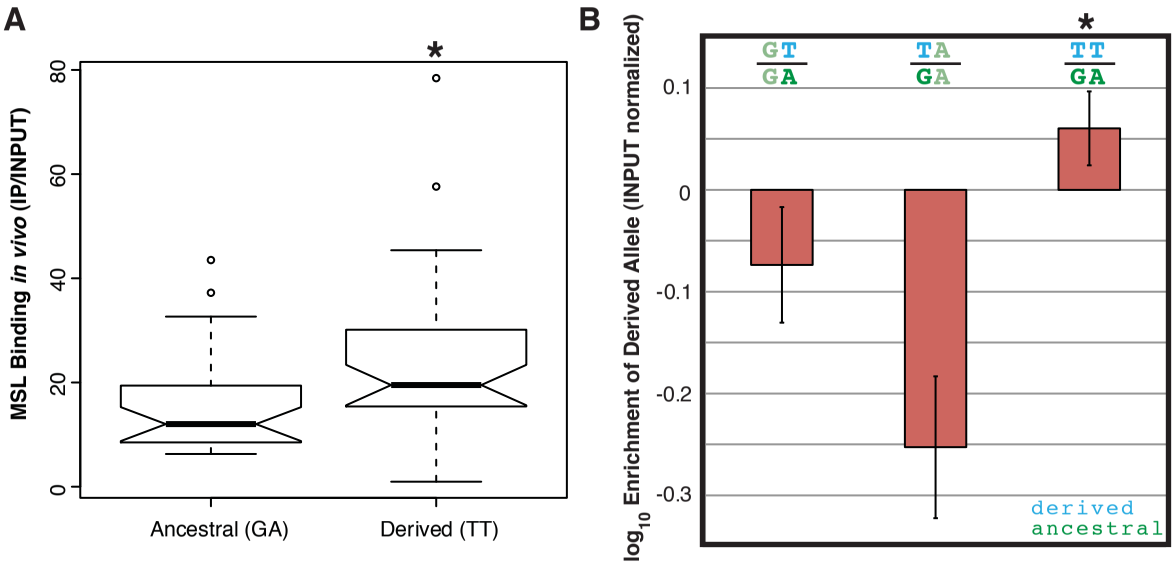
Figure 5 Selection shapes patterns of variation at the TT haplotype.

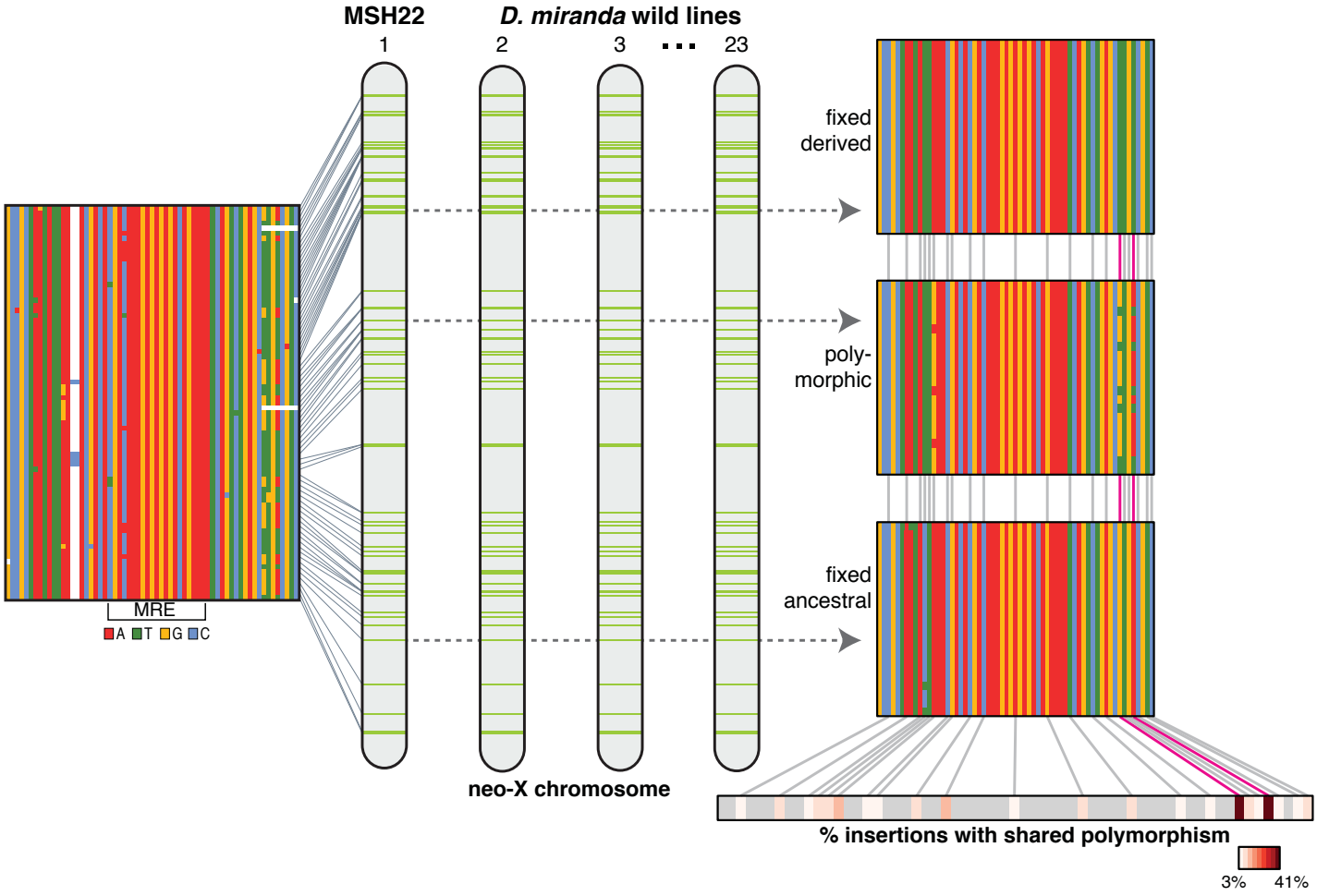
(A) Haplotype diversity across all ISX sequences. Assembled ISX contigs were combined for all insertions and individuals. The 25 basepairs flanking each side of the TT region were extracted from a total of 1,291 sequences and split into two groups based on whether they contained the TT or GA haplotype. Haplotype diversity was then calculated for each group. The difference between groups is significantly larger than expected by chance (resampling $P<0.001$), with the sequences containing the TT haplotype having less haplotype diversity compared to those containing the GA haplotype. (B) Nucleotide diversity across all ISX sequences. We compared nucleotide diversity for ISX insertions where all individuals carried the ancestral GA haplotype to those where the derived TT haplotype was fixed. ISX insertions that are fixed for the TT haplotype have significantly reduced nucleotide diversity compared to insertions fixed for the GA haplotype (one-sided Wilcoxon test P=0.035). (C) Allele-frequency spectrum across ISX sequences. The allele frequency spectrum was calculated separately for TT and GA-carrying ISX sequences, across all insertions and individuals. Consistent with incomplete hitchhiking under positive selection, the TT frequency spectrum shows an excess of high frequency derived alleles, compared to the GA spectrum (resampling $P=0.027$)..

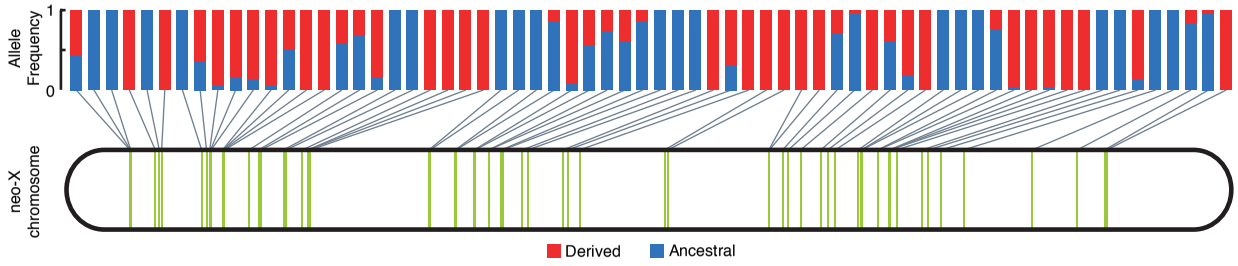
Figure 6. Non-allelic gene conversion spreads refining mutations among TE-derived MRE motifs

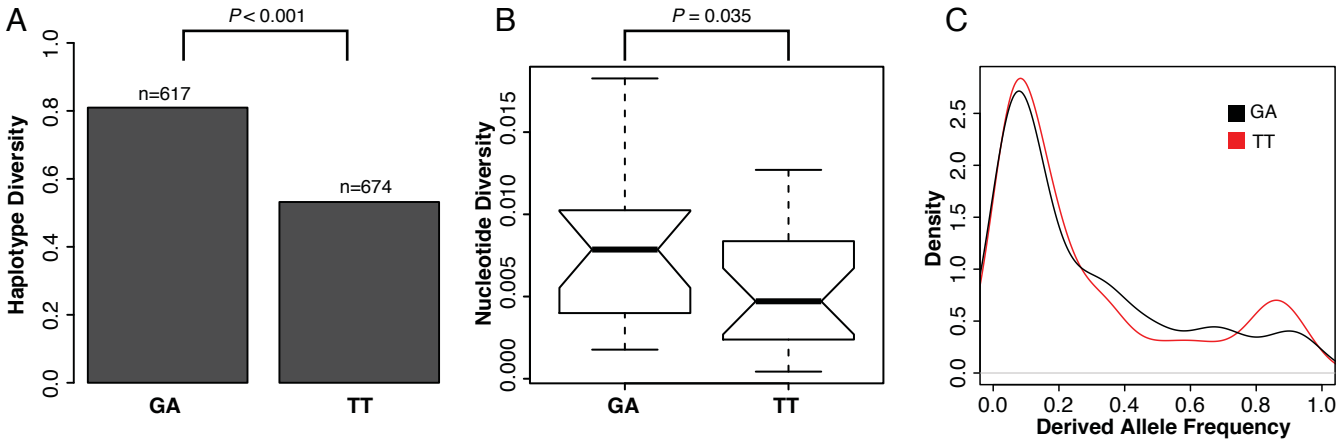
Shared polymorphism of the TT haplotype among ISX insertions suggests a model where a mutation that refines regulatory activity arose once at a single TE-derived regulatory element, and spread across elements via non-allelic gene conversion. Over evolutionary time, such a mutation spreads in two dimensions: horizontally among TE-derived regulatory elements and vertically through the population, until it is fixed across elements and across individuals. The TT haplotype is at the midpoint of this process. Across ISX insertions, it is fixed, absent, and polymorphic, in approximately equal proportions.

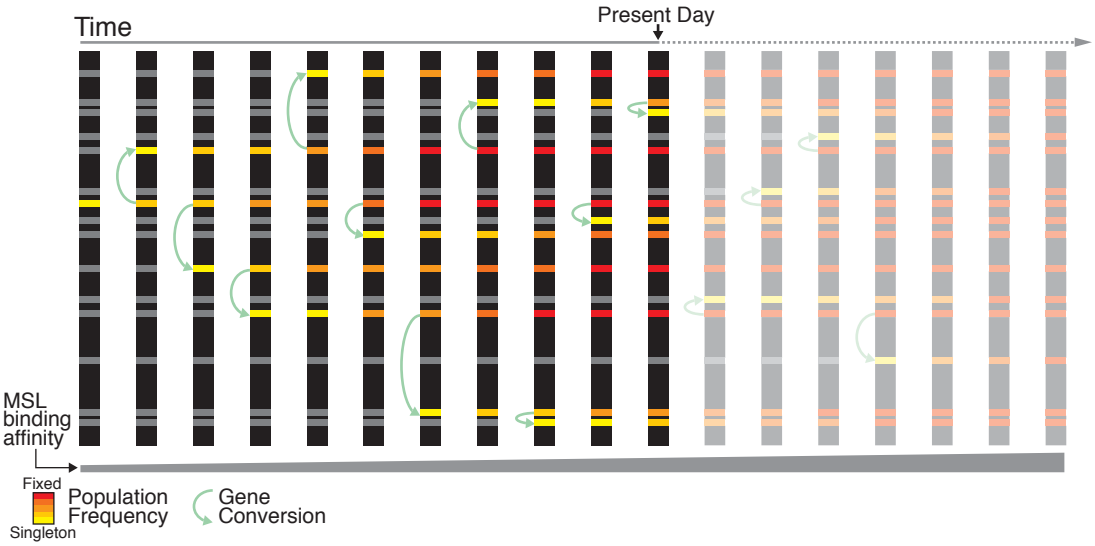

A
Derived Allele Frequency
0.0
0.2
0.4
0.6
ISX
ISY
MRE
A
T
G
C
B
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
ISX
ISY
GA
TT
GT
AA
TA
GC
CA
Haplotype


A
MSL Binding in vivo (IP/INPUT)
0
20
40
60
80
*
Ancestral (GA)
Derived (TT)
B
log10 Enrichment of Derived Allele (INPUT normalized)
0.1
0
-0.1
-0.2
-0.3
GT
GA
TA
GA
TT
GA
*
derived
ancestral


MSH22
D. miranda wild lines
1
2
3
23
fixed derived
poly-morphic
fixed ancestral
MRE
A T G C
neo-X chromosome
% insertions with shared polymorphism
3%
41%


Allele Frequency
1
0
neo-X chromosome
Derived
Ancestral


A
P < 0.001
n=617
n=674
Haplotype Diversity
0.0
0.2
0.4
0.6
0.8
1.0
GA
TT
B
P = 0.035
Nucleotide Diversity
0.000
0.005
0.010
0.015
GA
TT
C
GA
TT
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Derived Allele Frequency


Time
Present Day
MSL
binding
affinity
Fixed
Singleton
Population
Frequency
Gene
Conversion