EVOLUTIONARY BIOLOGY

The three-dimensional genome drives the evolution of asymmetric gene duplicates via enhancer capture-divergence

UnJin Lee[1,2]*†, Deanna Arsala[1]†, Shengqian Xia[1]†, Cong Li[2], Mujahid Ali[3], Nicolas Svetec[2], Christopher B. Langer[2], Débora R. Sobreira[4], Ittai Eres[4], Dylan Sosa[1], Jianhai Chen[1], Li Zhang[5], Patrick Reilly[6], Alexander Guzzetta[7], J.J. Emerson[8], Peter Andolfatto[9], Qi Zhou[3,10], Li Zhao[2], Manyuan Long[1]*

Previous evolutionary models of duplicate gene evolution have overlooked the pivotal role of genome architecture. Here, we show that proximity-based regulatory recruitment by distally duplicated genes is an efficient mechanism for modulating tissue-specific production of preexisting proteins. By leveraging genomic asymmetries, we performed a coexpression analysis on *Drosophila melanogaster* tissue data to show the generality of enhancer capture-divergence (ECD) as a significant evolutionary driver of asymmetric, distally duplicated genes. We use the recently evolved gene *HP6/Umbrea* as an example of the ECD process. By assaying genome-wide chromosomal conformations in multiple *Drosophila* species, we show that *HP6/Umbrea* was inserted near a preexisting, long-distance three-dimensional genomic interaction. We then use this data to identify a newly found enhancer (*FLEE1*), buried within the coding region of the highly conserved, essential gene *MFS18*, that likely neofunctionalized *HP6/Umbrea*. Last, we demonstrate ancestral transcriptional coregulation of *HP6/Umbrea*'s future insertion site, illustrating how enhancer capture provides a highly evolvable, one-step solution to Ohno's dilemma.

INTRODUCTION

Newly duplicated genes are at risk of loss in a population through genetic drift or negative selection (*1*) before they can acquire rare, advantageous mutations that lead to neofunctionalization. In *Drosophila melanogaster*, the probability of fixation for a slightly deleterious duplicate is one to two orders of magnitude lower than that of a neutral mutant ($P_{fix} = 0.085\sim0.003 \times 1/(2N_e)$, $N_e \approx 10^6$ as the effective population) (*1*, *2*). As a result, most nonfixed duplicate gene copies are expected to be lost within 2.32 generations or less (see Materials and Methods). These observations give rise to a longstanding evolutionary problem, referred to as "Ohno's dilemma" (*3*): Given the mutation rate in this species and the short time frame before loss, it is nearly impossible for a newly duplicated gene to reach fixation or acquire new mutations, especially the advantageous ones (*2*, *4*). Various models have been proposed with this problem including the duplication, divergence, complementation (DDC)/subfunctionalization model (*5*), the escape from adaptive conflict (EAC) model (*6*), and the innovation, amplification, and divergence (IAD) model (*3*, *7*).

The DDC model, also known as subfunctionalization, represents an evolutionary process where symmetric (identical) gene duplicates lose different aspects of their original function due to genetic drift. This random divergence results in the preservation of the duplicated genes, each retaining distinct, yet complementary, functions (Fig. 1, A and B). In contrast, genes evolving under the EAC model are under selection for enhanced optimization of specific functions originally held by the parental gene, which are then partitioned to paralogous copies. The EAC model posits that a single parental gene experiences intrinsic genetic conflict due to its inability to optimize multiple functions simultaneously, and gene duplication can resolve this evolutionary constraint (Fig. 1C). While the DDC and EAC models explain how ancestral functions are partitioned among gene duplicates, they do not adequately address the immediate development of altered expression patterns following gene duplication (Fig. 1, B and C). These altered expression profiles, often resulting from the gene's new genomic context, can be instrumental in driving the evolution of new functions—processes not fully captured by the DDC and EAC models. On the other hand, the IAD model (Fig. 1D) describes how shifts in selection pressures can promote the expression of genes with auxiliary functions by increasing gene copy number (Fig. 1). Following the initial increase of auxiliary function through gene amplification, subsequent relaxation of selection pressure allows changes to accumulate on the various copies, enabling the new copies to diverge and potentially gain a new function (*3*).

Although the IAD model provides a reasonable explanation for gene family expansions in microbial organisms while encountering environmental changes (*7*), it faces serious problems when applied to metazoans. A broad and general increase in gene dosage may be advantageous in some cell or tissue types but potentially deleterious in others (*8*, *9*). Like the EAC and DDC models, the IAD model does not directly explain how altered expression patterns arise immediately following gene duplication, leaving a gap in our understanding of duplicate gene evolution.

[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. [2]Laboratory of Evolutionary Genetics and Genomics, Rockefeller University, New York, NY, USA. [3]Department of Neuroscience and Developmental Biology, University of Vienna, Vienna, Austria. [4]Department of Human Genetics, University of Chicago, Chicago, IL, USA. [5]Chinese Institute for Brain Research, Beijing, China. [6]Department of Anthropology, Yale University, New Haven, CT, USA. [7]Department of Pathology, University of Chicago, Chicago, IL, USA. [8]Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA, USA. [9]Department of Biological Sciences, Columbia University, New York, NY, USA. [10]MOE Laboratory of Biosystems Homeostasis and Protection Life Sciences Institute, Zhejiang University, Hangzhou, Zhejiang, China.
*Corresponding author. Email: ulee@rockefeller.edu (U.L.); mlong@uchicago.edu (M.L.)
†These authors contributed equally to this work.

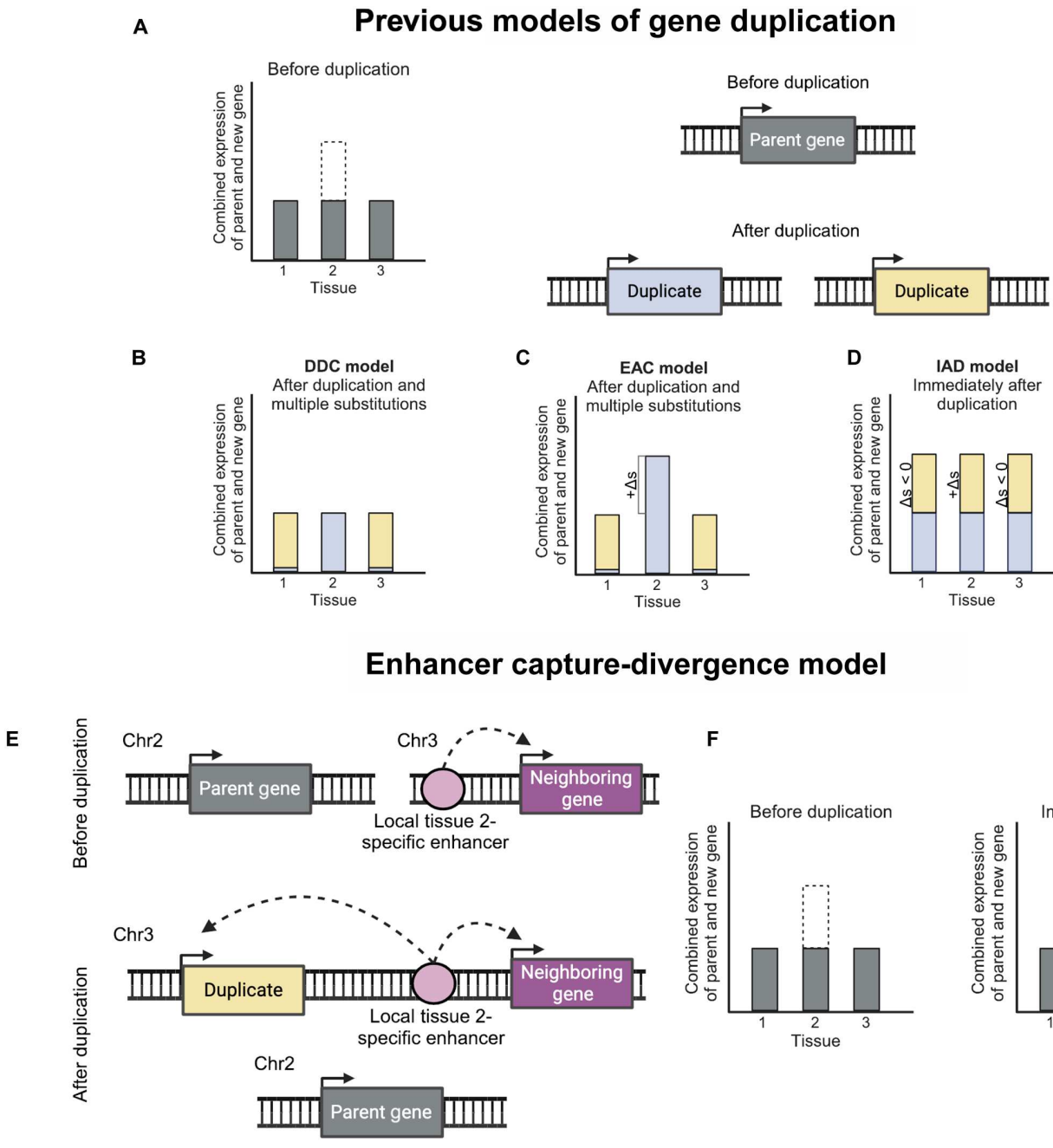


Fig. 1. Comparison of previous gene duplication models and the ECD model. (**A**) A preexisting gene that expresses in three tissues duplicates in the genome. Here, we assume that a selective advantage is conferred for increased expression in a single tissue, while all duplicate gene copies produce identical proteins. Dotted box represents selection for increased expression, and Δs indicates the change in selection coefficient. (**B**) In the DDC model, redundancy allows for compensation of any single loss-of-function event, eventually causing the expression pattern of the ancestral gene to be segregated between both new gene copies in a complementary fashion. As the total output of duplicate gene copies is identical to the original gene, the DDC model is a neutrally evolving process. (**C**) In the EAC model, internal conflict prevents increased tissue-specific protein production. This conflict is resolved via the act of duplication, where functions are segregated between duplicate gene copies, allowing the output of these two genes to increase fitness. (**D**) Under the IAD model, an ancestral gene duplicates, increasing production of the original protein in a single step. This increased dosage may cause deleterious effects via misexpression/overactivity in other tissues. Note that the identity of duplicate gene copies cannot be distinguished in the DDC, EAC, and IAD models (symmetric), resulting in random segregation of function or redundancy. (**E** and **F**) Under the ECD model, a parental gene fully duplicates into a distant region of the genome controlled by a preexisting enhancer. (E) By capturing this new interaction, this duplication increases tissue-specific production of the original protein in a single step. (F) Notice that the clearly identifiable parental gene copy remains unaltered, and thus all original function is retained, while the duplicate copy increases protein expression in a single tissue.

To address this gap, we propose the enhancer capture-divergence (ECD) model, which is an evolutionary model produced by asymmetric RNA or DNA-based gene duplication processes that allow for distinct parental and new gene identities and functions (Fig. 1, E and F). The ECD model first proposes that selective pressures change for the increased expression of a preexisting (parental) gene within a specific tissue or set of tissues (Fig. 1F). While the evolution of a new enhancer in the parental gene's locus is plausible, it would require multiple neutral de novo substitutions or insertions to generate one or more necessary transcription factor–binding sites that fix within a population and modulate the expression of the new gene duplicate without disrupting the parent gene's expression pattern. Under the ECD model, duplication of the parental gene into another regulatory environment under the control of a preexisting, tissue-specific enhancer is a solution that requires far fewer genomic changes and can occur in a single step. As the new selection pressures recur, the duplicate copy under new regulatory control will increase in frequency in the population, allowing it to become fixed. If the selection pressures change such that the increased tissue-specific expression of the new gene is no longer advantageous or if compensatory mutations appear in the original parent locus, selective pressures will relax on the new gene copy, allowing for divergence. While loss of the new gene copy by drift or negative selection is one possible fate, if the duplicate gene copy is at high enough frequency within a population, substitutions may accumulate and result in the gain of new, tissue-specific protein function.

The previous models addressing Ohno's dilemma (DDC, EAC, and IAD) are symmetric models of duplication-based evolution, which assume that the original parental gene function is randomly partitioned or entirely retained between identical duplicate copies, making parent and new gene copies indistinguishable from one another. They offer plausible mechanisms for the retention of certain types of gene duplicates in various processes of subfunctionalization, conflict resolution, and amplification of ancestral gene functions. In contrast, the ECD offers a more efficient route to neofunctionalization through the three-dimensional (3D) genome architecture rather than the development of new enhancers from scratch. It also provides a coherent and testable framework that describes the evolution of asymmetric or distal gene duplicates, which is currently unexplained by previous models. The asymmetry is a key feature of the ECD model that distinguishes it from the DDC, EAC, and IAC models in different consequences of functional evolution and allows for clear identification of genes that evolved under enhancer capture. (An extended discussion of prior models and genomic symmetry is available in the Supplementary Materials).

RESULTS

Analysis of tissue coexpression reveals that new genes evolve by enhancer capture

Central to the IAD model is the observation that gene duplication via unequal crossing over is more likely to occur than a point mutation (*3*, *7*). As previously described, one issue with this model is the implicit assumption that during the environmental shift, the increase in fitness gained by overactivity of the auxiliary function must be greater than the decrease in fitness imparted by overactivity of the original gene's function(s). In the case of the enzymatic activity of single-celled organisms where environments are encountered sequentially, it is reasonable to assume that selection might tolerate overactivity of the gene's original function during the transient environment in which the auxiliary function is favored. However, the decrease in fitness for improper expression or activity is larger in multicellular organisms than in single-celled organisms, where a multicellular organism's overall phenotype is the cumulative (development) and simultaneous (organ systems) product of many different gene functions.

In the case of multicellular organisms, selection may increase the expression of a gene within a single tissue type (Fig. 1). Under the IAD model, a full duplication of the parent gene function and expression pattern drives the duplicate copies to fixation as it provides the most evolvable solution to new conditions. In contrast, under the ECD model, a copy of the parent gene duplicates into a region of the genome containing an active enhancer(s) that modulates the new gene copy's expression in a tissue-specific manner. Alternatively, the new gene may duplicate into an inactive region of the genome containing unbound transcription factor–binding sites, thus activating a previously inert noncoding sequence, or a proto-enhancer, into an active enhancer.

Compared to the tissue-specific nature of genes evolving under the ECD model, genes evolving under the IAD model are overexpressed in all tissues, as they are assumed to take on the parent gene expression pattern. We therefore predict that enhancer capture will be more dominant than the IAD model for asymmetrically duplicated genes within multicellular organisms, as it avoids the potentially deleterious effects of increased dosage in multiple tissues resulting from full duplication (*8*, *9*). However, we stress that the IAD model is likely to drive the evolution of a large number of tandem duplicates as well as a subset of asymmetrically duplicates where the recruitment of preexisting regulatory elements is unlikely. This increase in fitness caused by the combined output of the new and parental genes thus drives the new gene copy to fixation, providing an alternate resolution to Ohno's dilemma than the IAD model. Once the tissue-specific selection for the new gene is relaxed, the new gene may then begin to diverge, accumulating substitutions.

Some classes of new genes will continue to evolve under the IAD, DDC, and EAC models. However, the relationships of new genes with their parent genes and neighboring genes differ in expression between those evolving under those previous models and our ECD model, allowing for direct testing of the ECD mechanism as a driver of newly evolved genes. Under the DDC or EAC models, the tissue expression patterns of parental and new genes are complementary, resulting in low coexpression between parental and new gene copies ("parental coexpression"). Since new gene evolution under the DDC and EAC models is assumed to occur in a regulatory-independent context, the tissue expression patterns of the new gene and its neighboring genes should have no relationship, resulting in random coexpression between the new gene and its neighboring gene ("neighboring coexpression"). Under the IAD model, genes and their upstream regulatory sequences are fully duplicated, which predicts a high coexpression between the parent and new gene copies, while the new gene copy and its neighboring genes should have low coexpression. In the ECD model, the parent gene is predicted to be more broadly expressed, while the new gene, which resides in a distant region of the genome, is under the control of one or more tissue-specific enhancers. Here, parental genes are expected to have broad tissue expression patterns, while new genes have expression patterns with high tissue specificity, resulting in low parental coexpression. On the other hand, since the new gene becomes

regulated by a locally captured enhancer that is already influencing other genes, neighboring coexpression is high, particularly in gene-dense genomes.

Tissue coexpression reveals that new genes evolve by ECD

We investigated the generality of the ECD model by first examining the coexpression patterns of newly duplicated genes and their parents in *D. melanogaster* ($N = 156$), including those that arose by tandem, distal, or retro-transposition–based duplication (*10*). We also used a separate, publicly available FlyBase dataset (*11*) containing 30 classes of developmental tissues produced by the modENCODE consortium (*12*) spanning 0- to 2-hour embryos to 30-day adults. Using these two data sources, we then calculated the Spearman correlation coefficient for the gene expression each new gene/parent gene pair across all 30 developmental conditions ("developmental coexpression"). Comparison of the developmental coexpression for tandem duplicates versus non-tandem duplicates (i.e., both distal duplicates and retro-transpositions) revealed significantly lower developmental coexpression in non-tandem duplicates than tandem duplicates ($P = 3.45 \times 10^{-10}$; fig. S2). When these comparisons were done with tandem duplicates versus either distal duplicates or retro-transpositions alone, non-tandem duplicates continued to show significantly lower developmental coexpression (distal: $P = 8.99 \times 10^{-9}$, retro-transposition: $P = 5.41 \times 10^{-3}$; fig. S2). This strongly suggest that the genomic location and type of duplication are critical in determining its expression pattern, demonstrating that regulatory neofunctionalization is a strong driver for non-tandem duplicates, as predicted by the asymmetric ECD model but not the symmetric DDC, EAC, or IAD models. Similar results have been observed in a wide range of studies, including but not limited to studies of retrotransposons and transposable element domestication as reviewed in (*13*, *14*).

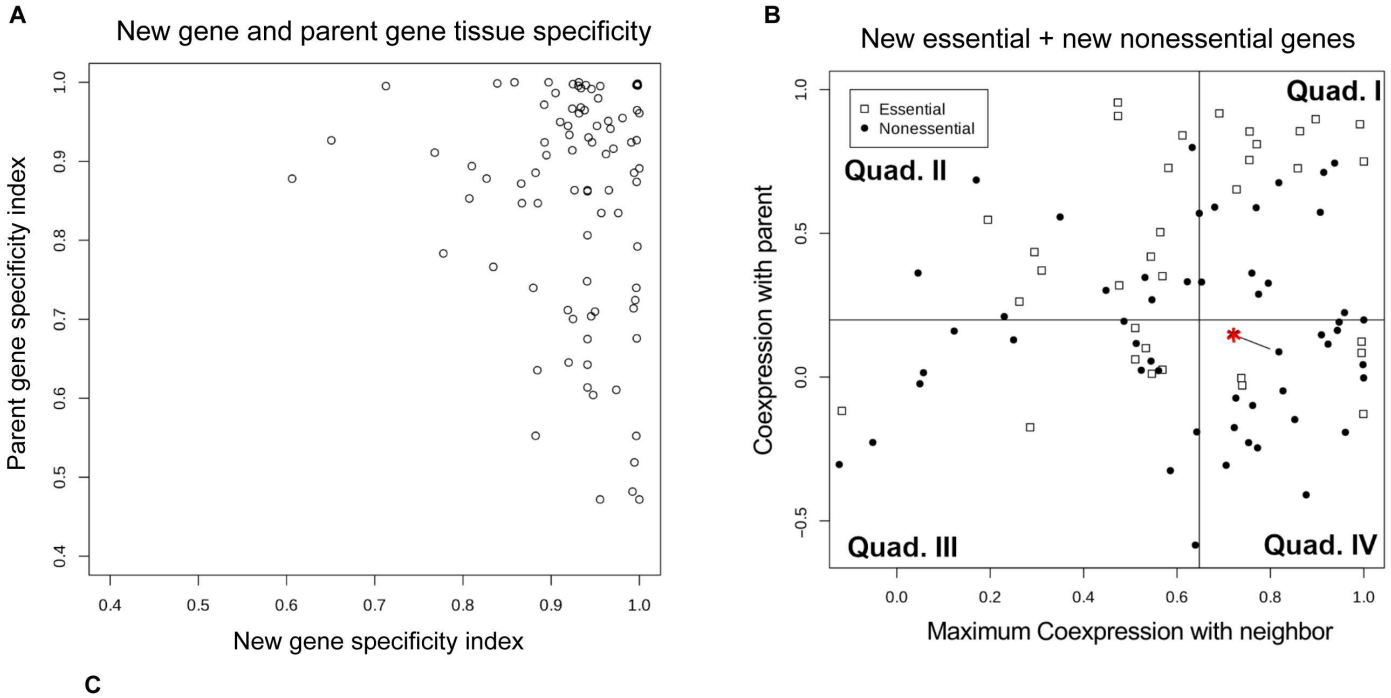
To determine whether the ECD process is a significant driver of new gene evolution in *D. melanogaster*, we obtained tissue expression data from FlyBase (*11*, *15*) (see Materials and Methods) and calculated coexpression between new/parental and new/neighboring gene pairs (Spearman correlation coefficient) for a random subset of newly evolved genes that (i) underwent a duplication into a different topologically associating domain (TAD) than its parental gene [as defined in (*16*)] and (ii) whose essentiality has been validated experimentally ($N = 87$; table S1 and Materials and Methods). We focused on experimentally validated genes that were in a different TAD, using distal duplications as a proxy, as their asymmetry allowed us to definitively identify the parent and new gene copies via synteny. To build a comprehensive assessment of gene coexpression patterns, we used expression data that contained tissue types extracted from both L3 larvae, prepupae, and adult flies, including gut, salivary glands, and imaginal discs from wandering L3 larvae, as well as the head, ovaries, gut, and reproductive organs from adults (see Materials and Methods). For tissues that were represented with multiple experimental runs, data from those tissue types were averaged before further analyses to avoid representation bias.

To determine whether a significant number of distally duplicated (non-tandem) genes evolved by enhancer capture, we used two concurrent features of the new genes in our dataset [parent/neighbor tissue coexpression ("PNC") and essentiality] that together determine whether the ECD process is a significant driver of new gene evolution alongside other established models (Fig. 2B). We define "low" and "high" coexpression as being below or above the median coexpression value across new genes in our dataset. Under the symmetrical DDC, EAC, and IAD models, the parent and new gene copies are indistinguishable in that all segregable and essential functions of the original gene partition randomly between both parent and duplicate gene copies. In contrast, the ECD model predicts that all original functions, including essential function, are expected to remain with the parental gene, while the new copy retains an auxiliary nonessential function. Thus, genes that evolved under enhancer capture are expected to be disproportionately enriched for nonessential functions. Furthermore, as genes under ECD are duplicated into a different regulatory environment from that of their parents, they are expected to appear in the lower-right quadrant (quadrant IV) in the PNC plots, with high neighboring coexpression and low parental coexpression (Fig. 2, B and C). In contrast, genes that have evolved symmetrically via the DDC or EAC models are expected to have a random partitioning of all functions (including essential functions) and should appear in the bottom half of the PNC plots (quadrants III and IV). Specifically, genes evolving under these processes are expected to have low parental coexpression resulting from divergent and complementary expression patterns, while the absence of regulatory context in the DDC and EAC models result in a prediction of random neighboring coexpression, as there is no expected relationship between the new gene and its neighboring genes (Fig. 2, B and C). Similarly, genes that have evolved via the IAD model should also have a random partitioning of essential functions while also appearing in the upper half of the PNC plots (quadrants I and II), with high parental coexpression resulting from full duplication (Fig. 2, B and C).

Alternatively, the ECD model predicts that most function, including essential gene function, will remain with the parental gene copy, while the tissue-specific expression pattern of the duplicate gene copy serves only to augment the function of the parental gene—a pattern frequently seen in new genes evolving via distal duplication (Fig. 2A). Specifically, selection for increased tissue-specific expression of the parental gene predicts the appearance of a distal duplicate of the parental gene copy both with nonessential function and high neighboring coexpression. Meanwhile, the expression pattern and gene function—including all essential function—of the parental copy remains unaltered and is retained. Together, the ECD model predicts a combination of high neighboring coexpression, low parental coexpression, and nonessentiality. This prediction may be tested by looking for a statistical enrichment of nonessential genes in the lower-right quadrant of the PNC plot relative to background (Fig. 2B and table S1). A distortion in the segregation of essential function is readily identified using the parent/neighbor coexpression plots for distally duplicated genes in *D. melanogaster*, where the ratio of new essential:new nonessential genes in the lower-right quadrant (5:17, 22.7%) was found to be significantly lower than the ratio of remaining new essential:new nonessential genes (29:36, 44.6%) [2.18-fold enrichment, $P = 0.0294$ binomial, $P = 0.0055$, 2D Kolmogorov-Smirnov (K-S) test based on coexpression data without median thresholding (*17*, *18*)], showing that enhancer capture is a significant driver of new gene evolution alongside previously established processes (Fig. 2).

HP6/Umbrea: A case study of a distally duplicated gene likely evolving under ECD

The evolution of the *HP6/Umbrea* locus provides an excellent example of the ECD model, as *HP6/Umbrea* is one of the few recently



C

Model	Parent/new gene coexpression	New/neighboring gene coexpression	PNC plot	Segregation of essential function
DDC	Low	Random	Lower half	Random
EAC	Low	Random	Lower half	Random
IAD	High	Random	Upper half	Random
Enhancer capture (*)	Low	High	Lower right	Parental gene

Fig. 2. Asymmetrically duplicated new genes evolve via enhancer capture. (**A**) Using new-gene/parent-gene pairs for genes evolving via distal duplication in *D. melanogaster*, the tissue specificity index τ is calculated and plotted above, demonstrating that new genes evolving via distal duplication have higher tissue specificity than parental genes. (**B**) Shown are PNC patterns for new genes in *D. melanogaster*, which have duplicated either more than 500 kb away or between chromosomes. Tissue coexpression (Spearman correlation coefficient) between new gene/parental gene pairs is plotted on the vertical axis, while maximal tissue coexpression between new gene/neighboring genes pairs is plotted on the horizontal axis. Vertical and horizontal lines indicate median coexpression value of all distally duplicated new genes presented here. Genes that evolved via enhancer capture are expected to have low parental coexpression and high neighboring coexpression and should thus be present in the lower-right quadrant. (**C**) While a new gene's essential function is equally likely to be partitioned between either parent or new gene under prior models, new genes evolving via enhancer capture are unlikely to have essential functions, as the expression of the new gene will only augment existing expression of the parental gene, leaving the original essential function intact. Comparing the ratio of new essential to new nonessential genes (29:36, 44.6%) in quadrants I to III to the ratio of new essential to new nonessential genes in quadrant IV showing high neighboring/low parental coexpression (5:17, 22.7%) shows that new genes likely evolve via regulatory capture (2.18-fold enrichment, $P = 0.0055$, 2D K-S test). (* denotes *HP6/Umbrea*).

evolved genes in *D. melanogaster*, whose protein evolution has been previously described in the literature [Fig. 2B, denoted as (*)] (*19*). Approximately 12 to 15 million years ago (Ma), *HP1b*, a gene located on the X chromosome, fully duplicated into a gene-poor, intronic region of *dumpy*, located on chromosome 2L (Figs. 3 and 4A). The duplication resulted in the new gene, *HP6/Umbrea*, which initially contained three known domains of its parent gene: the chromo domain, the chromo-shadow domain, and the hinge domain connecting the two.

HP6/Umbrea was lost ancestrally during multiple speciation events, suggesting that its original function may have been nonessential (*19*). This is consistent with a recent report that the extant form *HP6/Umbrea* is nonessential (*20*). *HP6/Umbrea* continued to evolve in a stepwise manner in the *melanogaster* lineage and rapidly diverged from its parental gene, *HP1b*. Following fixation, *HP6/Umbrea* lost its chromo domain approximately 10 to 12 Ma, which was then followed by sequence divergence and an accumulation of key substitutions 0 to 7 Ma, resulting in *HP6/Umbrea's* known centromeric protein function in *D. melanogaster* (*19*, *21*–*23*). The subsequent protein evolution after fixation suggests that protein neofunctionalization was not the driving force behind *HP6/Umbrea*'s fixation. Similarly, subfunctionalization and/or subsequent optimization of protein function can also be ruled out. This leaves enhancer capture as the only viable mechanism that may explain the fixation of *HP6/Umbrea*.

To determine whether enhancer capture drove *HP6/Umbrea*'s fixation, we examined the tissue expression of both *HP6/Umbrea* and *HP1b*. A simple comparison of *HP6/Umbrea*'s tissue-specific

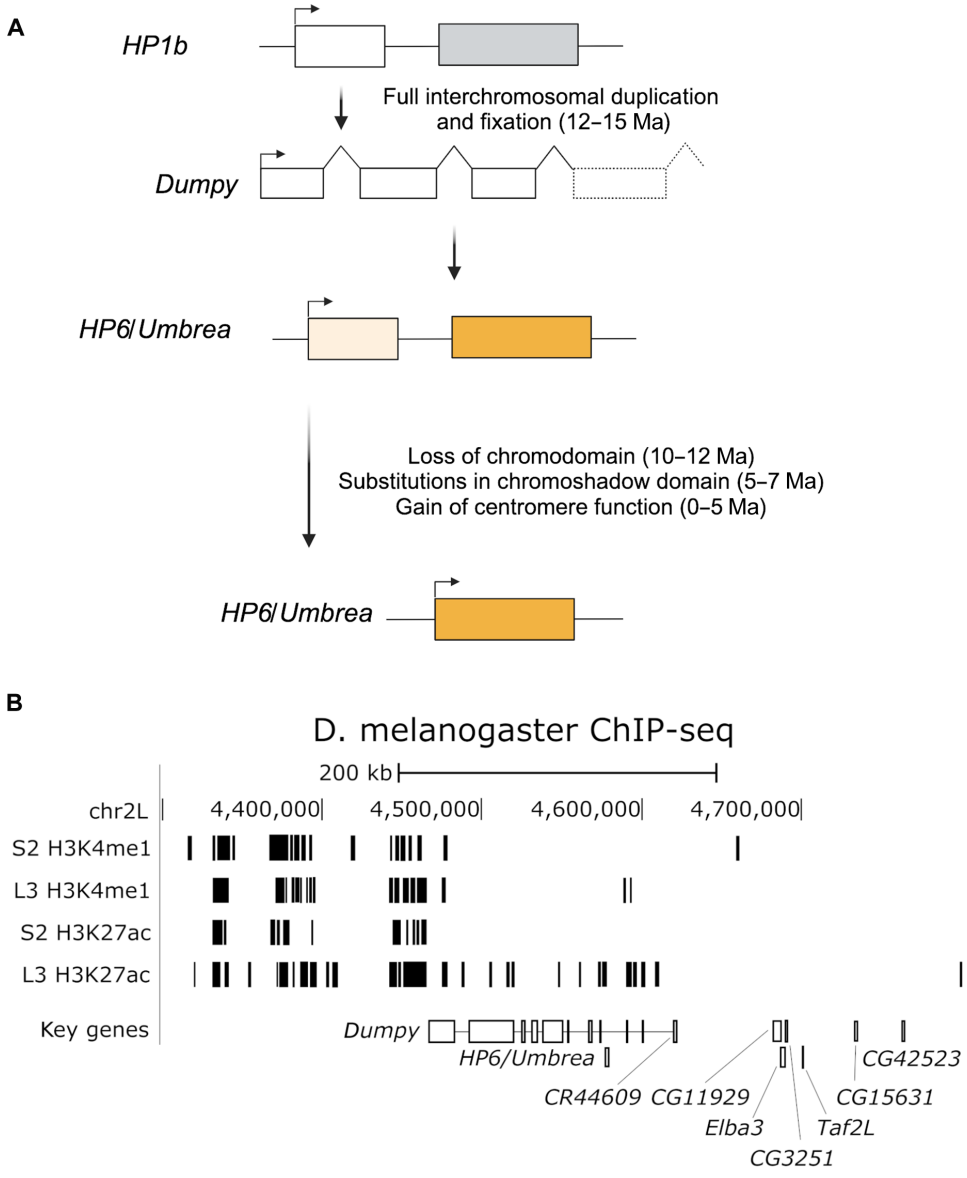


Fig. 3. HP6/Umbrea likely evolved via enhancer capture. (**A**) *HP6/Umbrea* is a new gene in *D. melanogaster* that arose from a full duplication of *HP1b* into an intronic region of *dumpy*, migrating from chromosome X to 2L. *HP6/Umbrea*'s well-characterized, step-wise protein evolution suggests that amino acid substitutions were unlikely to have driven the duplicate gene copy to fixation. (**B**) A comparison of ChIP-seq/ChIP-Chip markers for primed (H3K4me1) and active (H3K27ac) enhancers between embryonic S2 (no/low *HP6/Umbrea* expression) and whole L3 larvae (high *HP6/Umbrea* expression) tracks shows strong activation of a larval enhancer in a 100-kb intronic region of *dumpy* that is, aside from *HP6/Umbrea*, devoid of protein coding genes.

expression pattern to the parental gene *HP1b*'s very broad expression pattern suggests that *HP1b* is under constitutive regulation (fig. S1). Conversely, *HP6/Umbrea* is found only in a subset of tissues that express *HP1b*, suggesting that the new duplicate is under the control of one or more tissue-specific enhancers. *HP6/Umbrea*'s expression pattern is not similar to its first neighboring gene, *dumpy*, but its second neighboring gene, *CR44609*, which expresses in the imaginal discs, larval salivary glands, and male reproductive organs, which suggests that these genes are likely coregulated. The noncomplementary nature of the tissue expression patterns of *HP1b* and *HP6/Umbrea* provide further evidence ruling out subfunctionalization and/or subsequent optimization of regulatory function. These findings instead support the hypothesis that *HP6/Umbrea*'s expression is driven by the capture of tissue-specific enhancers, rather than the partitioning of ancestral regulatory elements.

Chromatin immunoprecipitation sequencing (ChIP-seq)/ChIP-Chip data (*24*) provide additional evidence that enhancer capture likely drove its early evolution. Compared to the embryonic S2 line in which *HP6/Umbrea* expression is minimal, we saw strong enhancer activity (denoted by active, H3K27ac, and primed, H3K4me1, histone marks) in whole L3 larvae within an intronic, gene-poor region of *dumpy*, coinciding with the onset of *HP6/Umbrea* transcription and its coregulated, neighboring gene, *CR44609* (Fig. 3B, compare histone activity in S2 to L3). Given the absence of other

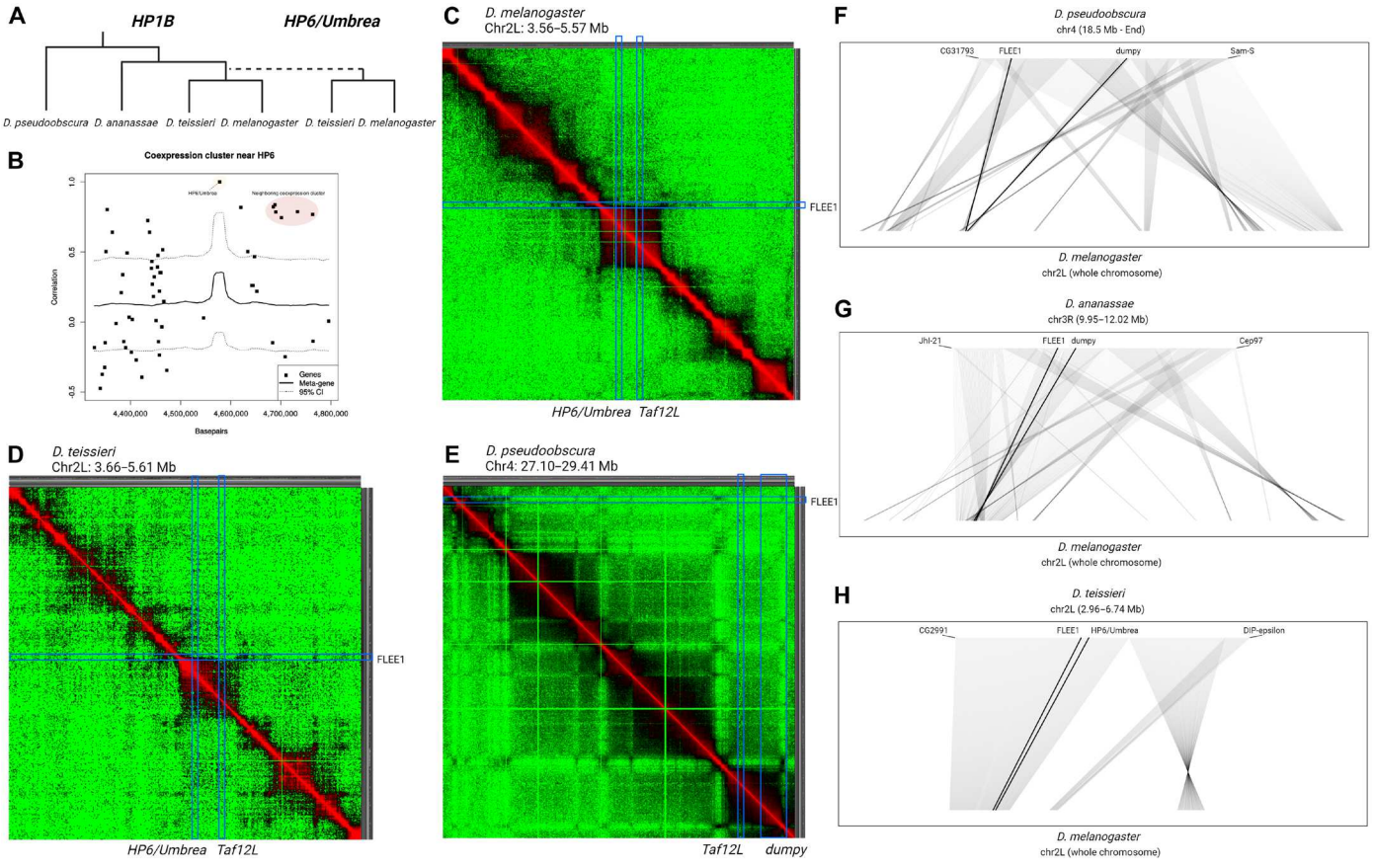


Fig. 4. *HP6/Umbrea* coexpression is associated with conserved chromosomal looping that predates its insertion. (**A**) Two in-group species, *D. melanogaster* and *D. teissieri* (div. ~6 Ma), contain *HP6/Umbrea*, while two out-group species, *D. pseudoobscura* and *D. ananassae* (pse-mel div. ~25 Ma, pse-ana div. ~12 Ma), predate *HP6/Umbrea*'s insertion (~12 Ma). (**B**) Tissue coexpression analysis between *HP6/Umbrea* and neighboring genes reveals the presence of a coregulated cluster of six neighboring genes. Note absence of other genes within *dumpy*'s intronic regions. (**C** to **E**) ~2-Mb Hi-C structures of (C) *D. melanogaster* chromosome 2L, (D) *D. teissieri* chromosome 2L, and (E) *D. pseudoobscura* chromosome 4 derived from whole adults and plotted at 1-kb resolution reveal the presence of multiple physically interacting compartments (green, low contact; black, medium contact; red, high contact). The chromosomal conformations of *D. melanogaster* and *D. teissieri* are broadly conserved but differ from *D. pseudoobscura*. The locations of *HP6/Umbrea* and/or *dumpy* (*HP6/Umbrea*'s future insertion site), the putative larval enhancer (*FLEE1*), and *Taf12L*, a member of the six-gene cluster, are notated. (**F** to **H**) Synteny between *D. melanogaster* chromosome 2L and (F) *D. pseudoobscura* chromosome 4, (G) *D. ananassae* chromosome 3R, and (H) *D. teissieri* chromosome 2L are assessed using gene order. Each line represents a single gene whose coordinates were lifted over between the respective chromosomes. The locations of *FLEE1*, *HP6/Umbrea*, and/or *dumpy* are shown in bold lines.

genes in the region (Figs. 3B and 4B), the presence of active enhancer marks in the region surrounding *HP6/Umbrea* provides compelling evidence for the role of enhancer capture in the regulation of this new gene.

It is far more likely that *HP6/Umbrea* duplicated into a region that appears to be under the control of a preexisting enhancer, rather than the independent emergence of multiple regulatory elements with similar tissue specificity in the vicinity of *HP6/Umbrea* and its coexpressed neighbors, which is unlikely given the short evolutionary time frame since its duplication. We tested for further coregulation in the region by applying a correlational analysis on bulk anatomical tissue expression data (see the section "Tissue coexpression reveals that new genes evolve by ECD") to determine whether *HP6/Umbrea* may be coregulated with other neighboring genes. We took a 500-kb region of the genome centered on the insertion site of *HP6/Umbrea* and calculated the tissue coexpression of each gene within this region in relation to *HP6/Umbrea*. As enhancers function in a proximity-based manner, we would expect a distance-dependent effect on the coexpression of neighboring genes across the genome. To generate a baseline estimate of this distance dependent coexpression distribution, we sampled 1000 random genic loci within the *D. melanogaster* genome, calculating the degree of coregulation expected on proximity alone. Notably, we found that using this distribution, the region of influence of any given regulatory region of the genome appears to be on the order of 25 kb, suggesting that this is a characteristic distance (1/e reduction) for enhancer interaction in *D. melanogaster* (Fig. 4B). Outside of this region of influence, the likelihood of coexpression relaxes to the genomic average. Therefore, genes found within this region of influence with high tissue coexpression with neighboring genes are potentially the result of coregulation with the focal gene. As expected, we find that the neighboring gene, *CR44609*, has a similar expression pattern as *HP6/Umbrea*. We also find that a locus of six neighboring genes (*CG11929*, *Elba3*, *CG3251*, *Taf12L*, *CG15631*, and *CG42523*),

which we refer to as the six-gene cluster, located approximately 100 kb away from *HP6/Umbrea* also expresses in the same tissues as *HP6/Umbrea*, expressing primarily in the imaginal discs, larval salivary glands, and adult male reproductive organs (Fig. 4A). The coexpression of *HP6/Umbrea* with its neighboring genes, both proximal and distal, suggests that the genomic region into which it was inserted is under the control of shared regulatory elements, supporting the enhancer capture model of evolution.

HP6/Umbrea and the six-gene cluster are contained within the same TAD

While *CR44609* and *HP6/Umbrea* coexpression may be explained simply because of their close proximity, the coexpression of the distal six-gene cluster is not immediately evident of coregulation. While the six-gene cluster is far outside of the *HP6/Umbrea* 25-kb region of influence, it may be physically located near to *HP6/Umbrea* because of the 3D nature of the genome and thus be coregulated. Similarly, the enhancer marks we previously identified located proximally to *HP6/Umbrea* (Fig. 3B) are also outside of the 25-kb region of influence but could still be nearby in a 3D space. To assess whether *HP6/Umbrea*, the putative enhancer, and the six-gene cluster physically interact, we analyzed publicly available Hi-C data in *D. melanogaster* derived from whole adults (Fig. 4C). A visualization of the local chromosomal conformations of the ~2-Mb region surrounding *HP6/Umbrea* reveals the presence of multiple locally interacting chromosomal domains. *HP6/Umbrea* and the six-gene cluster are contained within a clearly visible ~350-kb TAD (Fig. 4C).

Despite sharing a local chromosomal environment, the high frequency of shared physical contact is not necessarily indicative of shared regulation. To wit, most of the genes contained within the larger ~350-kb TAD do not share an anatomical tissue expression pattern with *HP6/Umbrea*. Thus, to determine whether *HP6/Umbrea* shows enriched physical contact with the neighboring six-gene cluster, we generated a 4C-Seq library from imaginal disc tissue derived from ~400 dissected L3 larvae and prepupae. This library was subsequently amplified using the *HP6/Umbrea* locus as a bait sequence. Visualization of the reads mapped from this library reveal a higher-resolution picture of the physical interactions between *HP6/Umbrea* and other loci contained within the same ~350-kb TAD (Fig. 4C and fig. S3). These data confirm the presence of a large degree of physical interaction between the *HP6/Umbrea* locus and the six-gene cluster. The significance of this interaction is further demonstrated by comparing our 4C-Seq peaks to a virtual 4C analysis derived from Hi-C data (Fig. 5A). A Circos visualization (*25*) of statistically significant Hi-C interactions [$P < 0.0001$, to reduce number of depicted interactions, significance calculated in HOMER (*26*)] anchored on the *HP6/Umbrea* locus shows broad concordance between our 4C-Seq peaks and those identified by Homer (*26*).

The *HP6/Umbrea* locus structure is likely driven by a tissue-specific larval enhancer

Our 4C-Seq analyses also revealed the presence of a several putative regions with enriched contact (fig. S3), including revealing highly enriched contact with a single, distal 394–base pair (bp) locus located roughly 130 kb away from the *HP6/Umbrea* locus. This locus was expanded by approximately 750 bp on both 5′ and 3′ ends and was named the Four-C Larval Enhancer Element (*FLEE1*) (data S1). The 2165-bp *FLEE1* construct was found to be entirely contained within the coding regions of the genes *MFS18* and *Elp3*, which are both highly conserved, essential genes in *D. melanogaster*.

To validate whether *FLEE1* contained a functional larval enhancer, we assayed pGreenRabbit reporter plasmids, which we site-specifically integrated in *D. melanogaster* (Bloomington *Drosophila* Stock Center 79604) (*27*). Compared to control homozygote transformants that contain empty reporter vectors driving basal levels of green fluorescent protein (GFP) expression, we found that *FLEE1* directed GFP expression in the salivary glands of third instar larvae (Fig. 5, E to J, and movie S1). This result is consistent with prior in vivo studies demonstrating *HP6/Umbrea*'s key role in the telomeres of polytene chromosomes in larval salivary glands (*28*).

FLEE1 is housed within the coding sequences and 3′ untranslated region of two protein-coding genes, *MFS18* and *Elp3* (fig. S4). Because *MFS18* and *Elp3* are essential genes, we were unable to perturb these sequences without introducing confounding effects. However, a population genetic analysis of the *MFS18* locus reveals that the coding sequence of *MFS18* is under selective pressure not only to maintain/conserve *MFS18* amino acid sequence but also to maintain regulatory function as an active larval enhancer. The *FLEE1* locus shows strong divergence from *Drosophila yakuba* and *Drosophila simulans* while maintaining low levels of polymorphism within natural populations in *D. melanogaster*, suggesting that the locus is under strong selective pressure (fig. S5). However, an analysis of the ratio of nonsynonymous (K_a) to synonymous (K_s) substitution rates from *Scaptodrosophila lebanonensis* to *D. melanogaster* for *MFS18* shows that the vast majority of these substitutions are synonymous substitutions ($K_a/K_s = 0.033$, $P = 0.0022$) (*29*), demonstrating that this locus is under strong purifying selection. Alternatively, the *MFS18* locus fails to show signatures of directional selection, being unable to show significance in the correct direction under the Hudson-Kreitman-Aguadé (HKA) (*30*) and McDonald-Kreitman (MK) tests (table S2) (*31*). These combined results suggest that the coding sequence of *MFS18* is under selective pressure to maintain/conserve *MFS18* amino acid sequence while simultaneously maintaining regulatory function as an active larval enhancer, displaying a stereotypically high substitution rate as is common with enhancers under stabilizing selection (*32*). These results stand in sharp contrast to the *HP6/Umbrea* locus, which shows signatures of strong directional selection under both the HKA and MK tests (table S2).

3D contact between larval enhancer and six-gene cluster predates chromosomal rearrangements and *HP6/Umbrea* insertion

While the functionally validated *FLEE1* enhancer demonstrated highly specific tissue expression, the ECD model predicts that this interaction predates the insertion of *HP6/Umbrea*. To determine whether this physical interaction is ancestral to *HP6/Umbrea* insertion, we used a cross-species comparison of whole adult-derived Hi-C data produced from *D. melanogaster* (Fig. 4C), *Drosophila teissieri* (*HP6/Umbrea* present, divergence ~6 Ma) (Fig. 4D), and *Drosophila pseudoobscura* (*HP6/Umbrea* absent, divergence ~25 Ma) (Fig. 4E). Broad conservation of the chromosomal conformation of *HP6/Umbrea* locus between *D. melanogaster* and *D. teissieri* demonstrates how the local 3D structure has remained relatively stable over the past 6 million years. In addition, the larval *FLEE1* enhancer, *HP6/Umbrea* and the six-gene cluster are present in the same TAD in both species. However, a visual comparison between

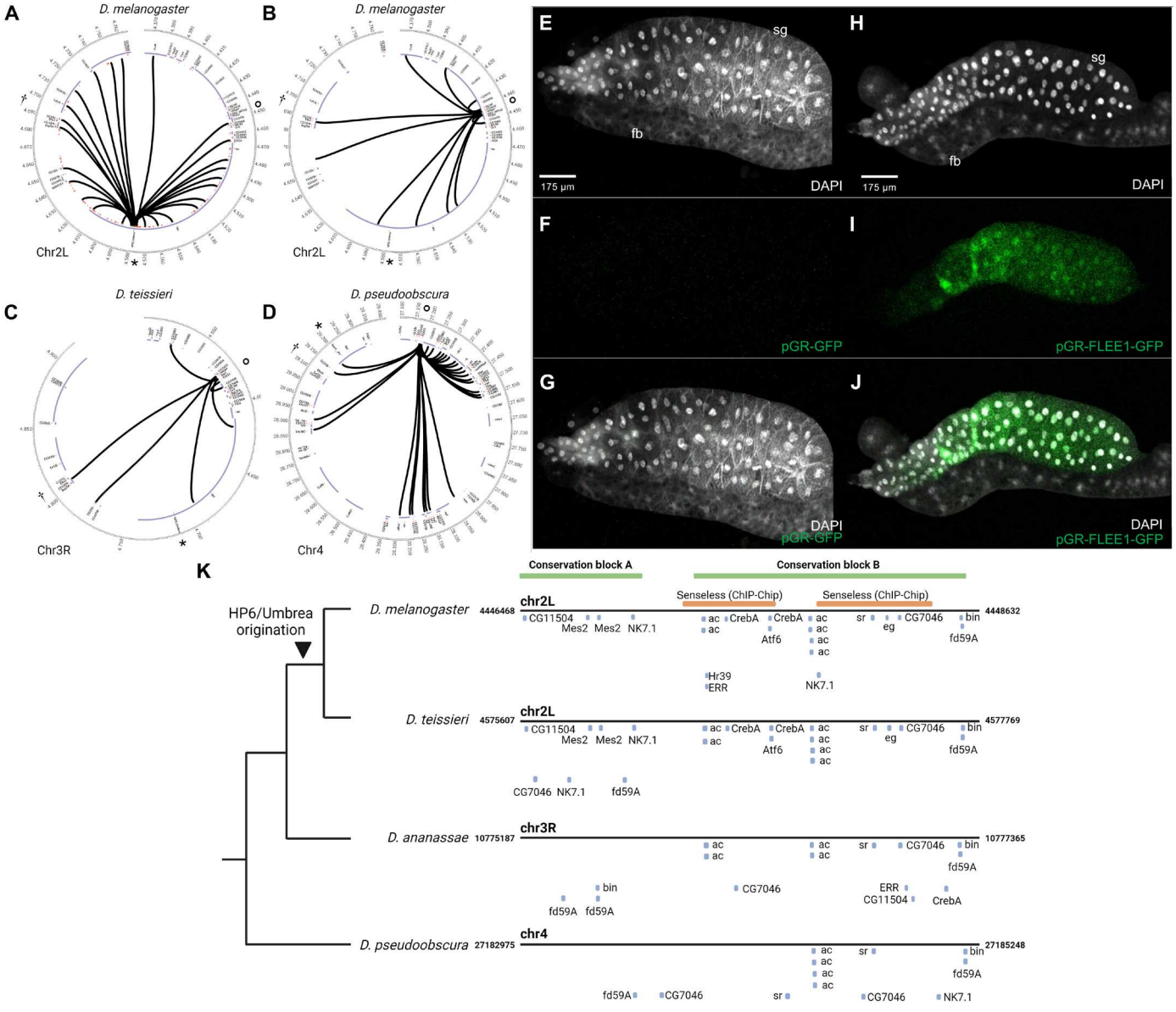


Fig. 5. *FLEE1* encodes a larval, salivary gland enhancer. (**A**) Circos diagram of virtual 4C track of the *HP6/Umbrea* (denoted by *) locus shows a large degree of contact with the six-gene cluster of coexpressing genes (denoted by †) as well as the putative larval enhancer (*FLEE1*, denoted by °). Gene locations are indicated by the blue track with respective labels. Nonzero coverage regions of a 4C-Seq experiment are also overlaid on the red track, showing broad overlap between virtual 4C contacts and regions identified by 4C-Seq. (**B** to **D**) Circos diagrams of virtual 4C tracks of the *HP6/Umbrea* locus centered on *FLEE1* are shown for (B) *D. melanogaster*, (C) *D. teissieri*, and (D) *D. pseudoobscura*, showing significant contact between *FLEE1* (denoted by °), the six-gene cluster (denoted by †), and either *HP6/Umbrea* or its future insertion site (denoted by *). (**E** to **J**) The *FLEE1* enhancer was tested for enhancer activity in third instar larvae using the pGreenRabbit reporter vector. [(E) to (G)] Basal GFP reporter expression from an empty reporter vector in a third instar salivary gland and fat body. [(H) to (J)] GFP reporter expression directed by *FLEE1* in the salivary gland, with minimal expression in the fat body. Green, GFP. White, 4′,6-diamidino-2-phenylindole (DAPI) (DNA). sg, salivary gland. fb, fat body. (**K**) Analysis of predicted transcription factor–binding sites on *FLEE1* shows that two blocks of sequence have been conserved between *D. melanogaster* and *D. pseudoobscura* (conservation block A on left; conservation block B on right). Conservation block A evolved at a similar time period as the origination of *HP6/Umbrea*, while conservation block B predates *HP6/Umbrea* origination. Note that two *CrebA* sites and one *Atf6* site in conservation block B also evolved around the same time as *HP6/Umbrea*. The location of two ChIP-Chip–derived *senseless*-binding sites in conservation block B is also shown.

D. melanogaster and *D. pseudoobscura* showed that the presence of *FLEE1* and the six-gene cluster in a shared TAD is not the ancestral state. While the linear distance between *FLEE1* and the six-gene cluster is around 250 kb in both *D. melanogaster* and *D. teissieri*, the same linear distance is approximately 2 Mb in *D. pseudoobscura*. This reduction of linear distance between *FLEE1* and the six-gene cluster resulted from a series of chromosomal rearrangement events that occurred between *D. pseudoobscura*, *Drosophila ananassae*, and *D. melanogaster* (Fig. 4, F to H).

Despite the 2-Mb separation between *FLEE1* and the six-gene cluster, examination of the Hi-C data demonstrated a large degree of physical interaction between the TADs containing *FLEE1* and the

six-gene cluster. However, a demonstration of inter-TAD interactions does not necessarily indicate significant interactions between *FLEE1* and the six-gene cluster. We thus performed a virtual 4C analysis using *FLEE1* as a viewpoint in *D. melanogaster*, *D. teissieri*, and *D. pseudoobscura* (Fig. 5, B to D, and Materials and Methods). As a positive control, we find that the virtual 4C track in *D. melanogaster* shows significant interaction [$P < 0.0001$ to reduce number of depicted interactions, significance calculated in HOMER (*26*)] between *FLEE1*, *HP6/Umbrea*, and the six-gene cluster as expected. An analysis of *D. teissieri* also reveals that these interactions remain significant [$P < 0.05$, HOMER (*26*)]. The physical interaction between *FLEE1*, the six-gene cluster, and the future insertion site of *HP6/Umbrea* (the first intron of *dumpy*) is not statistically significant in the virtual 4C track of *D. pseudoobscura* derived from whole adults (not shown). However, the same track using *D. pseudoobscura* data derived from L3 larvae (Fig. 5D and fig. S6) shows that the interaction between *FLEE1*, the six-gene cluster, and the future insertion site of *HP6/Umbrea* is statistically significant, suggesting that these interactions are conserved [$P < 0.05$, HOMER (*26*)]. A virtual 4C track in *D. pseudoobscura* larvae centered on the ancestral *HP6/Umbrea* insertion site further confirms that the interactions with the ancestral six-gene cluster and ancestral *FLEE1* locus predates *HP6/Umbrea* origination (fig. S7) [$p < 0.05$, HOMER (*26*)].

FLEE1 enhancer structure predates *HP6/Umbrea*

While the statistically significant, specific chromosomal interactions between *FLEE1* and the six-gene cluster suggests a large degree of functional importance, this does not necessarily indicate whether this conserved function is regulatory in nature or it is indicative of an ancestral regulatory role for the *FLEE1* larval enhancer. To determine whether *FLEE1* potentially had enhancer activity before *HP6/Umbrea* insertion, we performed a binding site analysis for known transcription factors using the *FLEE1* sequence for *D. melanogaster*, *D. teissieri*, *D. ananassae*, and *D. pseudoobscura* using the MEME Suite [FIMO (*33*)] and the CIS-BP 2.0 database (*34*). Note that only transcription factors with a *q* value of <0.05 in *D. melanogaster* were retained for these analyses.

The regulatory structure of *FLEE1* consists of two blocks of conserved transcription factor–binding sites (Fig. 5K). The first block, conservation block A, consists of *CG11504*, *Mes2*, and *NK7.1* sites, which have been conserved between *D. melanogaster* and *D. teissieri*. The second block, conservation block B, consists of *achaete*, *stripe*, *biniou*, and *fd59A* binding sites, which have been conserved between all four examined species. An analysis of modENCODE transcription factor–binding sites (*35*) in FlyBase (*36*) revealed the presence of only empirically validated transcription factor–binding sites within FLEE1 in the case of two *senseless* ChIP-Chip signals within Conservation Block B (Fig. 5K). The absence of *senseless*-binding sites in our FIMO analysis suggests that the appearance of these *senseless*-binding sites may be contingent on the presence of multiple cofactors rather than being strictly dependent on sequence. The organization of *senseless*-binding sites overlapping with or in close proximity to *achaete*-binding sites suggests that *achaete* may be this key factor driving *senseless* binding. This is consistent with prior functional studies demonstrating *senseless*' role as a directly bound coactivator of *achaete* (*37*). Despite the lack of ChIP data for *senseless* binding in non-*melanogaster* species, the coactivation of *senseless* and *achaete* provide a testable hypothesis for the functional conservation of *achete*-binding sites before *HP6/Umbrea* insertion.

Cell type–specific coexpression of *HP6/Umbrea*'s future neighboring genes predates its insertion

The conservation of 3D contact between the future *HP6/Umbrea* locus, the neighboring six-gene cluster, and distal functional elements does not necessarily imply regulatory conservation. Therefore, to determine whether *HP6/Umbrea* and its neighboring genes are coexpressed and whether those neighboring genes coexpress in species without *HP6/Umbrea*, we used single-cell RNA sequencing (scRNA-seq) in the following species: *D. melanogaster* and *D. yakuba*, both containing *HP6/Umbrea*, and *D. ananassae*, which diverged before the origination of *HP6/Umbrea*. Note that *D. yakuba* and *D. teissieri* are sister species that diverged approximately 1 Ma (*38*). We performed scRNA-seq in the testis tissue because of its high evolutionary importance (*23*, *39*–*41*), the existence of preexisting high-quality cell type annotations (*42*, *43*), and the higher expression levels of *HP6/Umbrea* and its coexpression cluster in this tissue type relative (*42*) to the imaginal disc or salivary gland tissue (fig. S1).

After mapping and visualization of the scRNA-seq data using previous cell type annotations (Fig. 6, A and B) (*42*, *43*) as well as data from all three species on the same, shared projections (Fig. 6, A and B, and fig. S8), it becomes clear that *HP6/Umbrea* is likely co-regulated on a cellular level with the entire coexpression cluster, while overall expression of *Elba3* and *CG3251* are low and restricted mainly to germline stem cells (GSC)/early spermatagonia, late spermatogonia, and early spermatocytes. As an internal control, somatic and developmental cell types cluster together as expected. The bulk of the expression is shared across the coregulated genes, while further cell type–specific expression is also shared within GSCs/early spermatogonia and late spermatogonia. *CG11929*, *Taf12L*, and *CG15631* all show shared cell type–specific coexpression in *D. ananassae*, demonstrating an ancestral coregulation of these genes. A more quantitative examination of these expression patterns reveals more subtle expression pattern differences (Fig. 6D). The significantly lower and cell type–specific expression of *Elba3* and *CG3251* remains evolutionary conserved across the three species. While *CG11929* and *CG15641* show peak expression in early spermatocytes, *Taf12L* shows peak expression in GSC/early spermatogonia and late spermatogonia. This early peaking of transcription is similar to that of *HP6/Umbrea*, suggesting strong coregulation of *HP6/Umbrea* and *Taf12L*. This early peaking behavior of *Taf12L* is conserved in *D. ananassae*, demonstrating how this expression pattern is ancestral to *HP6/Umbrea* origination.

DISCUSSION

Identification of distal larval enhancer

FLEE1's regulatory activity residing within the primarily exonic regions of the highly conserved *MFS18* gene constitutes an example of how protein-coding regions of the genome may also have key regulatory functions (*44*, *45*). Such pleiotropy demonstrates how the interpretation of synonymous substitution rates may not necessarily serve as good estimates of neutral evolution rates in commonly used codon table–based tests of molecular evolution. Rather, substitutions typically regarded as synonymous could alternatively be indicative of strong directional or stabilizing selection for the regulatory function of genomic enhancer elements. Furthermore, elucidation of the *FLEE1-HP6/Umbrea* interaction highlights the importance of identifying and characterizing the contributions of structural variations and chromosomal rearrangements in driving phenotypic

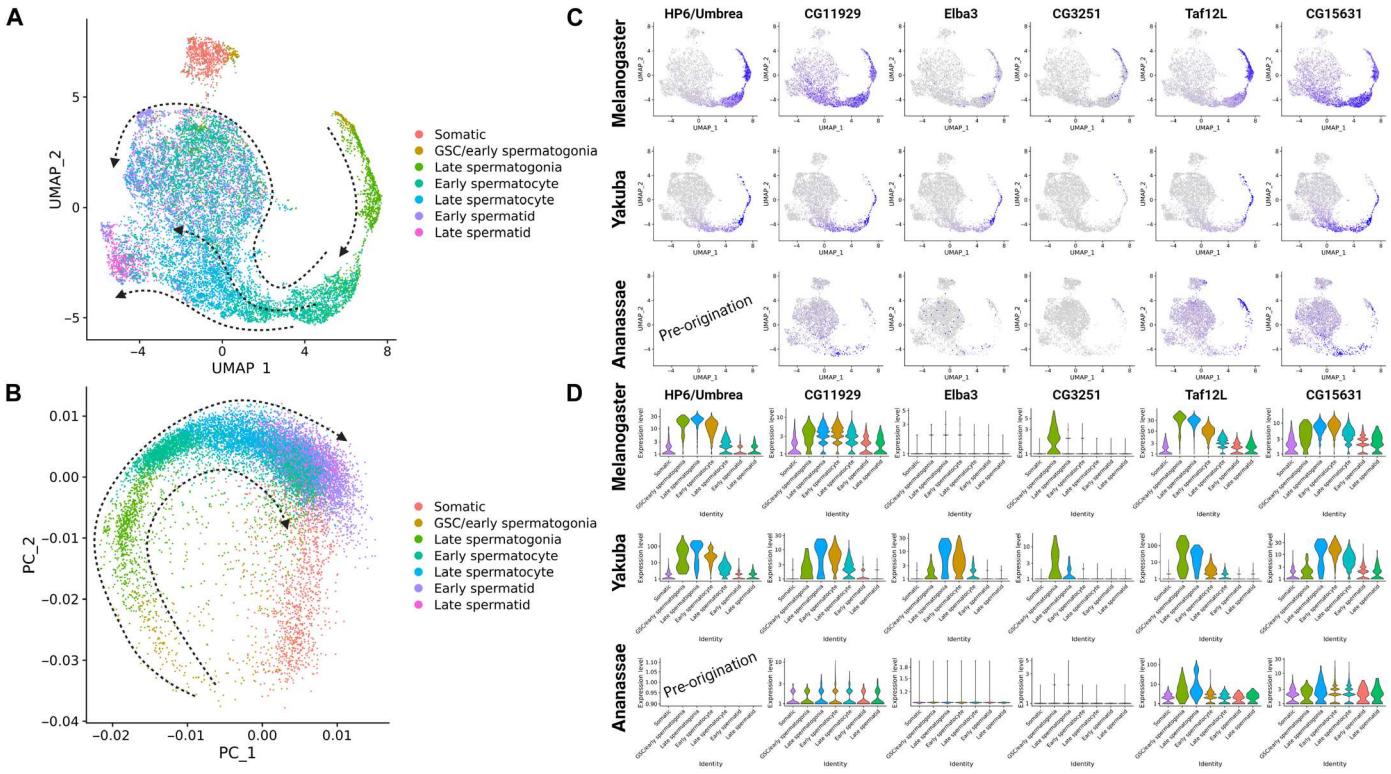


Fig. 6. Coregulation of the six gene cluster is ancestral. scRNA-seq data from *D. melanogaster*, *D. yakuba* (both containing *HP6/Umbrea*), and *D. ananassae* (*D. melanogaster–D. ananassae* divergence predating *HP6/Umbrea* origination) are mapped to the same manifold using a core 198-gene list. (**A**) Preexisting data from and their corresponding labels from prior studies were included in this Uniform Manifold Approximation and Projection, allowing for precise identification of somatic and developmental (GSC/early spermatogonia, late spermatogonia, early and late spermatocyte, and early and late spermatid) cell type clusters both within and across species. The developmental trajectory of spermatogenesis is indicated using lines and arrows. (**B**) A principal components analysis demonstrates the consistency of the scRNA-seq data, with PC1 corresponding to developmental time (stem-like states on the left and mature states on the right) and PC2 corresponding to germline or somatic cell states (somatic states on the bottom, germline states on the top). (**C**) Expression data from *HP6/Umbrea* and the coexpression cluster of genes are plotted for all three species (CG42523 excluded as it did not pass quality controls for all species in Seurat). Expression patterns (raw counts) for all depicted genes are heavily biased toward GSC, spermatogonia, and early spermatocytes. (**D**) Finer quantitative analysis of cell type–specific expression patterns using violin plots [log(raw counts+1)] demonstrates how *Taf12L* shows an identical expression pattern as *HP6/Umbrea*, where expression is highest in GSC/early spermatogonia and late spermatogonia cell types and drops off as spermatogenesis continues. This cell type bias is conserved in *D. ananassae*, predating the origination of *HP6/Umbrea*.

evolution (*46*). Similarly, while the molecular function of *HP6/Umbrea*'s neighboring coexpressing lncRNA long noncoding RNA gene, *CR44609*, remains unknown, it was identified as one of 170 newly evolved genes that may have evolved under the cultivator model of de novo gene origination (*43*). The authors demonstrate how the evolutionary activation of these genes alters the transcriptional bursting properties of neighboring genes during a key stage of spermatogenesis (*43*). These results suggest that the putative function of *CR44609* may be to regulate a neighboring cultivator gene (e.g., *HP6/Umbrea* or *dumpy*) in cis via short-range, chromosomal effects [e.g., supercoiling-mediated transcription coupling or promoter interference (*47*, *48*)], further highlighting a key role that chromosomal organization may play in phenotypic evolution.

Two *CrebA* sites, a leucine zipper transcription factor associated with regulation of tissue-specific genes in the salivary gland (*49*), and one *eagle* site were found in conservation block B for *FLEE1* and were found to be conserved between *D. melanogaster* and *D. teissieri*. Prior studies reveal that *CrebA* directly activates genes encoding core components of endoplasmic reticulum cargo translocator and *Cop I/II* vesicle secretory machinery (*50*, *51*) via binding of salivary gland polytene chromosomes (*52*). A secondary function of *CrebA*, along with *senseless*, is the maintenance of salivary gland–specific genes encoding for secreted cargo, transmembrane proteins, and enzymes (*51–53*). This observation is consistent with a prior study of polytene chromosomes demonstrating overlap between a subset of genomic *CrebA* and *senseless*-binding sites (*52*). In addition, *eagle* has been shown to function in a specific and combinatorial fashion with *Huckebein*, a key cell-fate specifier for salivary gland morphogenesis (*49*, *53*), to guide differentiation of serotonergic cells (*54*). While functional evidence from non-*melanogaster* species in the form of perturbative experiments and/or reporter assays remains lacking, the cis-regulatory logic of *FLEE1* provides evidence supporting a possible ancestral role of *FLEE1* in salivary gland expression. Alternatively, it is also possible that salivary gland expression of *FLEE1* is a secondary function of ancestral *FLEE1* before *HP6/Umbrea*. In that case, the concurrence of the origination of *HP6/Umbrea* and the appearance and stabilization of two conserved *CrebA* sites and conservation block A may be suggestive of a more bidirectional influence of cis- and trans-regulatory elements, where the linking of a particular protein function with an existing

enhancer may alter or stabilize that enhancer's regulatory function. In the case of *HP6/Umbrea*, the original recruitment and retention of the original *HP1b* duplicate is associated with changes in the ancestral *FLEE1* locus, where a more distal single *CrebA* site was lost while the two aforementioned *CrebA* sites appeared in closer proximity to the existing conserved *achaete*-binding sites. While questions remain regarding the evolutionary history of *FLEE1*, the likelihood of an ancestral tissue-specific function of *FLEE1* remains high as indicated by the conservation of multiple binding sites as well as its long-distance physical interactions despite multiple chromosomal rearrangement and speciation events.

Our results demonstrate how stabilizing selection for the conservation of large-scale chromosomal conformations may drive the appearance of evolutionary novelty resulting in the development of novel, centromeric function as in the case of *HP6/Umbrea*. Further work to investigate the regulatory role of the *FLEE1* enhancer will require tissue-specific epigenetic interference at its locus to assess the transcriptional impacts of its target genes. While additional work will be required to reveal what evolutionary forces underlie the strong conservation of the long-distance interaction of the *HP6/Umbrea* locus before the new gene's insertion, our findings demonstrate how complex chromosomal conformations are a key, underappreciated element in the evolution of the eukaryotic genome (*46*).

ECD model

ECD joins various previously proposed models to interpret different evolutionary aspects of gene functionality. Whereas DDC and EAC, for example, explain the duplication-dependent subfunctionalization and resolution of adaptive-caused conflict from ancestral genes with multiple functions respectively, ECD interprets neofunctionalization for creating evolutionarily novel gene functions through duplication. ECD demonstrates how the manner of duplication itself may provide neofunctionalization in an asymmetric, tissue-specific manner. Such neofunctionalization provides a selective advantage directly through the single-step acquisition of regulatory elements, a process made possible by the 3D organization of the genome, which brings distant enhancers into close spatial proximity to the duplicated gene. Evolutionary changes to genome organization over time can also create new opportunities for enhancer capture and drive the emergence of evolutionarily novel gene functions. These newly acquired regulatory elements for distally duplicated genes maintain the duplicate for an adequate amount of time until new, advantageous mutations occur that solve Ohno's dilemma.

In addition to partial duplication phenomena such as the generation of gene fusions (*55*) as well as favorable frameshifts (*56*), our model highlights the underappreciated evolutionary value of both the act of duplication itself and, more importantly, the genomic context in which these duplications occur. While the role of positional effects in gene regulation and evolution has long been appreciated (*46*, *57*, *58*), the advent of new chromosomal conformation capture technologies allows us to directly connect the conservation of chromosomal domains (*59*, *60*) and the origination of new genes under a strong conceptual framework.

Under the ECD model, a gene copy duplicates into a preexisting regulatory context (Fig. 7A), gaining a new regulatory interaction. This model thus provides a mechanistic explanation by which gene interaction networks may rapidly evolve (*40*). Under this model, we have two separate gene interaction subnetworks for both parental and neighboring genes (Fig. 7B). As a new gene duplicates into a region near the neighboring gene, the new gene acquires the upstream regulatory function of the neighboring gene as well as the original parental gene's downstream protein function (Fig. 7C) while simultaneously preserving the preexisting interactions from both parental and neighboring genes' subnetworks. Since duplication has been observed to occur more frequently than point mutations (*3*, *7*), enhancer capture provides a faster route to generating increased tissue-specific expression of a parental gene (Fig. 1) than any set of mutations in the parental gene's regulatory sequence. Duplication in the 3D looping architecture of the eukaryotic genome recombines genes and enhancers into new combinations, thus resulting in regulatory novelty (Fig. 7C). Hence, this model provides an explanation and mechanism for the well-described but poorly understood phenomenon where new gene duplicates often have highly tissue-specific expression patterns (Fig. 2A) (*41*, *61*, *62*).

One key aspect of the ECD model is the selective advantage imparted by increased tissue-specific expression. The resolution of genetic conflict, such as sexual antagonism, is becoming increasingly appreciated as a driver of the evolution of new genes (*63*, *64*). While most new genes have highly tissue-specific expression patterns, these often favor either the female or male reproductive organs/germ lines in *D. melanogaster* (*41*). A close examination of the expression pattern of *HP6/Umbrea* demonstrates the same—*HP6/Umbrea* is expressed primarily in the imaginal discs, larval salivary gland, and the male reproductive organs. Hence, it is possible that the selective advantage imparted by *HP6/Umbrea*'s original duplication may have been a result of regulatory sexual antagonism, and, given that most new genes show expression specific to reproductive organs, enhancer capture may be a widespread mechanism for the resolution of sexual antagonism. Furthermore, *HP6/Umbrea*'s repeated ancestral loss suggested that it was originally nonessential following duplication but later gained its semilethal phenotype in a step-wise manner (*19*). While questions remain regarding the specific molecular function of *HP6/Umbrea* (*20*), its interaction with *FLEE1*, and other genes located in quandrant IV (Fig. 2B), suggesting the need for further perturbative experiments, our results highlight the importance of the 3D genome structure for understanding new gene origination.

One question underlying the ECD model is what proportion of newly evolved genes originates through the ECD process. This question remains intrinsically difficult to answer for many reasons. It is impossible to determine the likeliest ancestral expression patterns for parent and new gene pairs in the absence of equivalent anatomical bulk RNA-seq datasets from out-group species. Even if full tissue-specific transcriptomes were made available for a large panel of closely related species, further complications would arise in interpretation of such data, as the DDC, EAC, IAD, and ECD processes are not mutually exclusive. Specifically, the expected tissue expression patterns for each process show a large degree of overlap, with the only differentiator between each process being the degree to which reproductive fitness changes as a result of different selective pressures. For example, DDC, EAC, and ECD processes could theoretically produce the same transcriptional signatures in parent and new gene pairs arising through entirely different selective mechanisms. Alternatively, newly evolved genes may not be limited to only one evolutionary process. For example, it is possible that a gene may undergo distal duplication to increase tissue-specific expression under the ECD. This could then be followed by tandem duplications that thus increase tissue-specific dosage under the IAD process, which

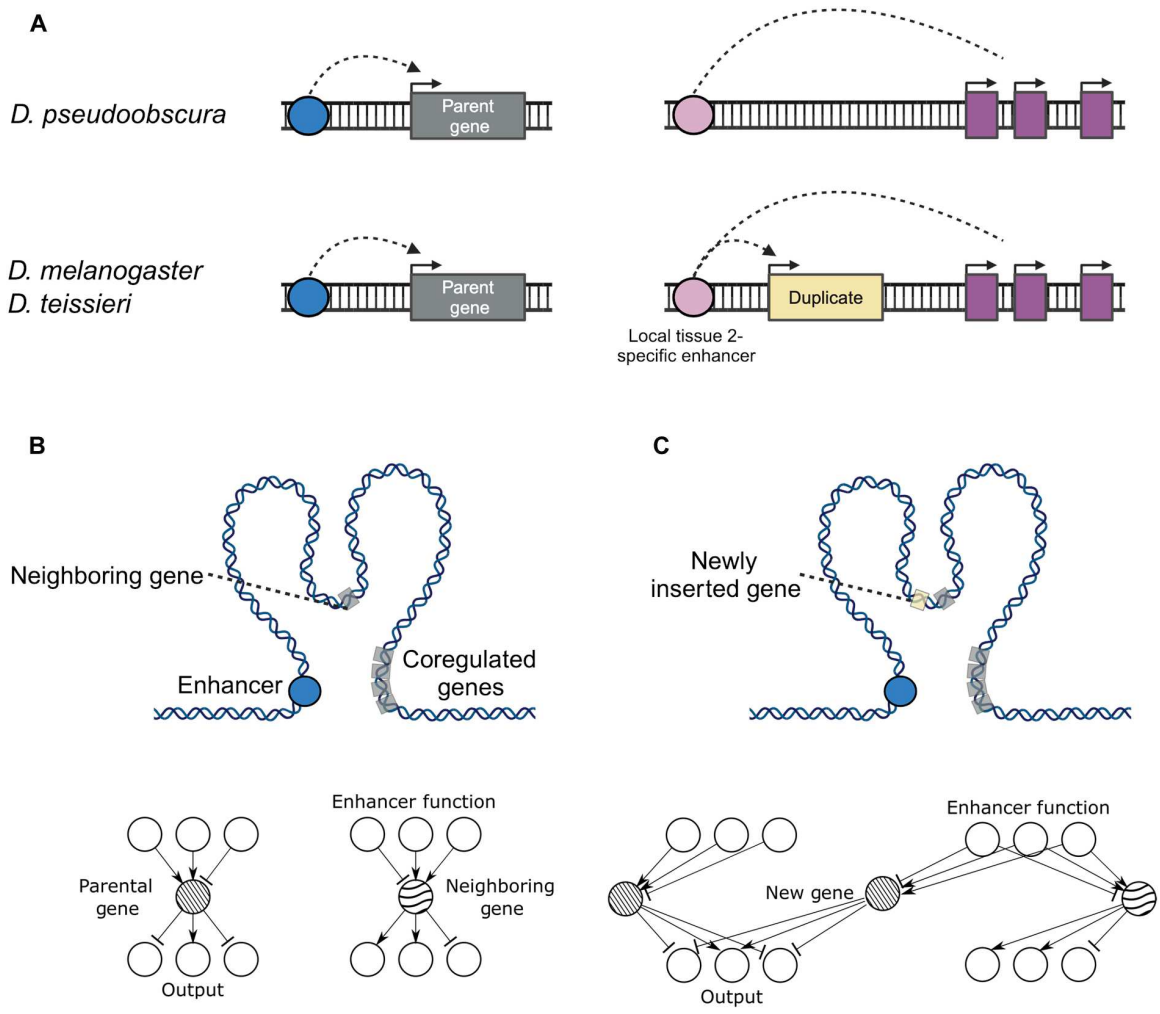


Fig. 7. The 3D organization of the genome allows for rapid rearrangement of genetic networks. (**A**) Depicts a cartoon illustration of the action of the larval enhancer on the neighboring cluster of coregulated genes as well as the future insertion site of *HP6/Umbrea*. Preceding insertion of *HP6/Umbrea*, the larval enhancer was in contact with both *HP6/Umbrea*'s neighboring gene as well as with the coregulated six-gene cluster. (**B** and **C**) This looping structure remains conserved following *HP6/Umbrea*'s insertion, allowing for a rapid recombination of elements upstream of *HP6/Umbrea*'s neighboring gene (i.e., larval enhancer) with elements downstream of *HP6/Umbrea*'s parental gene (i.e., *HP1b*'s protein function). A sample gene interaction network, both pre- and post-duplication, is also depicted above. Note that the parental gene and neighboring gene's original interactions remain intact, preserving previous function.

could subsequently allow for complementary degeneration and/or optimization of function under the DDC or EAC processes. Further complications arise when considering effects such as partial duplication or the appearance of chimeric sequences that still retain some partial function of the parental gene.

The methodology presented here is dependent on high tissue specificity as well as low levels of transcriptional divergence after initial selection. While it is possible that many genes in all quadrants of Fig. 2B evolved via ECD, we use the observation that genes in the lower-right quadrant are likeliest to have evolved under enhancer capture. However, despite our inability to precisely identify the specific evolutionary mechanism underlying the ancestral fixation of each new duplicate gene copy, we exploit key symmetries present in the DDC, EAC, and IAD models to be able to demonstrate the generality of the ECD process. A prior study of newly evolved species- and clade-specific genes also provides further indirect evidence for the generality of the ECD model by demonstrating that most functionally important duplicate gene copies arise via dispersed duplication and retro-transposition events (i.e., distal duplication events) (*10*). Specifically, newly evolved genes specific to either *D. melanogaster* or *D. yakuba* (i.e., species-specific genes) were compared to newly evolved genes that are shared between *D. melanogaster*, *D. simulans*, and *Drosophila sechellia* (i.e., clade-specific genes). While 81.9% (*D. melanogaster*) and 78% (*D. yakuba*) of species-specific genes arose through tandem duplication, only 33.9% of retained clade-specific genes evolved via tandem duplication. Conversely, only 15.2% (*D. melanogaster*) and 22.0% (*D. yakuba*) of species-specific genes evolved via distal duplication events, while 54.3% of clade-specific genes evolved via distal duplication events. This pattern of low survivorship of young tandem duplications suggests that these newly evolved distal duplicates likely have important but distinct (although not necessarily essential) function than parental genes, a key prediction of the ECD model.

The molecular mechanism of ECD is a general process

The molecular mechanism of enhancer capture, as described in the ECD model, is a general process that is not limited to evolution of new genes but has also been observed in other biological contexts, such as the development of human malignancies. The first found example of enhancer capture occurring in cancer is the t(8;14) translocation in Burkitt's lymphoma, allowing for the oncogene *Myc* to be expressed under the regulatory control of the *immunoglobulin heavy chain gene* (*IGH*), which is expressed in lymphoid cells (*65*, *66*). To date, there have been a variety of other examples of enhancer capture rearrangements involved in oncogenesis occurring in diverse tissues (*67*–*71*). Although most rearrangements bring oncogenes into proximity with constitutive regulatory elements of a given cell type, they may also be brought into proximity with context-specific regulatory regions. One such translocation in prostate cancer involves the translocation of the oncogenes *ETV1* or *ERG* within proximity of the promoter region of *TMPRSS2*, which contains several androgen receptor binding sites. In this instance, *ETV1* or *ERG* gains androgen-dependent expression, which can be abrogated by androgen deprivation therapy, a common treatment for prostate cancer (*72*, *73*).

One longstanding question in this literature is why particular rearrangements are commonly associated with specific cancers. Are common rearrangements observed because they are the few examples that confer a selective advantage in a select cell type or does the cancer cell type have a structural predisposition to favor those rearrangements? Although common rearrangements do confer a relative fitness advantage to the cancer cells, the evidence has become clear that recombining loci are likely to be within physical proximity of one another (*74*, *75*). It has been shown that chromosomes 9 and 22 neighbor each other in hematopoetic cells, which may explain the frequency of the t(9:22) translocation in chronic lymphocytic leukemia, which produces the *BCR-ABL* fusion protein (*76*, *77*). In addition, it has been shown that the *Myc* and *IGH* genes are brought within close physical proximity during B cell stimulation (*78*) highlighting the importance of cell context–specific genomic arrangements in cancer.

The primary difference between enhancer capture in cancer and organismal evolution is the lack of necessity for cancer cells to preserve an oncogene's previous function via gene duplication before translocation. In addition, cancer cells typically experience selection at the clonal level, so rearrangements do not need to confer optimized gene expression within multiple tissue contexts. However, the cancer literature is clear that enhancer capture is a commonly occurring one-step mechanism that allows individual cells to gain fitness advantages and that cell type–specific 3D genome confirmations selectively favor certain rearrangements. Given our finding that the ECD model is a significant driver of new gene evolution, it is likely that the inherent 3D configuration of the germline genome imposes an important and previously unappreciated constraint on evolutionary novelty.

MATERIALS AND METHODS

Tissue expression data and analysis

Tissue expression data were retrieved from FlyBase. Precomputed reads per kilobase million (RPKM) data files were downloaded, with RPKM values for each FlyBase transcript being reported for 29 tissues (*15*). As many of the tissue types were repetitive, data from the head, ovary, carcass, and digestive system were averaged to reduce overrepresentation bias in further correlational analyses. Gene map data were also obtained from FlyBase to properly identify neighboring genes (*11*). Parental/new gene pair information was retrieved from (*23*). Spearman correlation coefficients were calculated using the tissue expression data between parental and new gene pairs. Because of intronic structures and variation in gene length, two neighboring genes for each new gene on each side were assessed using Spearman correlation coefficients and the maximum value of the four neighbors was recorded. In addition, correlation coefficients for all genes within 500 kb of *HP6/Umbrea* were reported. To generate a baseline distance-dependent genomic estimate of coexpression, 1000 random genic loci were chosen and coexpression values (Spearman) between the randomly selected gene and all neighbors within a 500-kb range were calculated. This 500-kb region was then divided into 100 nonoverlapping windows, where means and variances in correlation coefficients were calculated across all randomly selected loci.

ChIP-seq data

ChIP-seq or ChIP-Chip data were obtained for H3K4me1 and H3K27ac for S2 cells as well as whole L3 larvae from modENCODE (*24*). H3K4me1 ChIP-Chip data for S2-DRSC cells were obtained using data ID 304 and 3760. H3K27ac ChIP-Chip data for S2-DRSC cells were obtained using data ID 296 and 3757. H3K4me1 ChIP-seq data for whole Oregon-R L3 larvae were obtained using data ID 4986. H3K27ac ChIP-seq data for whole Oregon-R L3 larvae were obtained using data ID 5084. For all datasets, data were obtained in .gff3 format and visualized using the UCSC Genome Browser.

Hi-C data

We generated Hi-C datasets for *D. pseudoobscura* using whole L3 larvae (SRR23968954, 32 Gbases) and used publicly available Hi-C datasets for whole adults derived from *D. melanogaster* (SRR5820094, 93.1 Gbases), *D. teissieri* (SRR12331760, 33.4 Gbases), and *D. pseudoobscura* (SRR11813284, 119.1 Gbases). Hi-C libraries were processed entirely using Homer (v5.0.1, dockerhub: avianalter/homer_hic) (*26*, *79*) and bowtie2 (v2.5.4, dockerhub: staphb/bowtie2) (*80*). Specifically, homerTools trim was used to trim reads to the appropriate restriction digest sites (see respective Sequence Read Archive entries), and reads were mapped using bowtie2. Reads were mapped to the latest RefSeq assemblies available for *D. melanogaster* (GCF_000001215.4, dm6), *D. teissieri* (GCF_016746235.2, Prin_Dtei_1.1), and *D. pseudoobscura* (GCF_009870125.1, UCI_Dpse_MV25). Because of computational limitations, downstream analyses were limited to the chromosomes containing *HP6/Umbrea* and/or *FLEE1* as determined by BLASTn (*D. melanogaster*: NT_033779.5, *D. teissieri*: NC_053029.1, *D. pseudoobscura*: NC_046681.1; see "Genome annotations"). Tag directories for each genome/Hi-C dataset combination were generating using Homer and subsequently processed using Homer's "analyzeHiC" software with 1-kb resolution and 5-kb super-resolution parameters for all datasets. Resulting raw interaction matrices were visualized using TreeView (v1.2.0) (*81*) as recommended in the Homer user manual. Matrices were $\log_2$ scaled within TreeView, and contrast/threshold settings were adjusted as needed following the Homer user manual. Significant chromosomal interactions were detected using Homer's analyzeHiC software using a resolution of 5 kb, a super-resolution of 10 kb, and a minimum interaction distance of 7.5 kb. Circos diagram configuration files were generated using Homer with genome annotations produced using

LiftOff (see "Genome annotations") (*82*) and visualized using Circos (v0.69-6, dockerhub: alexcoppe/circos) (*25*). Virtual 4C tracks were generated with the resulting Circos configuration files through the addition of a rule conditioning visualization to only links present on either 1-kb regions flanking *HP6/Umbrea* and/or *FLEE1*.

Genome annotations

Genomic sequences for chromosomes containing *FLEE1* and/or *HP6/Umbrea* using a command-line version of BLASTn (v.2.13.0+) (*83*). FASTA sequence files for each chromosome were extracted and used as target sequences for LiftOff (v1.6.3, dockerhub: avianalter/liftoff) (*82*) using the *D. melanogaster* chromosome 2L sequence and annotations as queries. Lifted over annotations were limited to "transcript," "exon," "CDS," "start_codon," and "stop_codon" fields. Synteny was assess using gene ordering via LiftOffTools (v.0.4.3, dockerhub:avianalter/liftofftools) (*84*) and visualized using R (v.4.0.2).

4C-seq data

About 400 *D. melanogaster* L3 larvae and prepupae were freshly dissected in 10-min intervals on ice. A single-cell suspension was generated from imaginal disc tissue using collagenase. These suspensions were pooled and formaldehyde-fixed for 10 min, followed by glycine quenching. Aliquots of these suspensions were quantified and snap frozen with liquid nitrogen and stored at −80°C until 10^7 cells were accumulated. All cells were then collected and resuspended in a lysis buffer containing Triton X-100, NP-40, and protease inhibitors followed by homogenization via douncing. Nuclei were then gently lysed using a SDS and Triton-X while shaking (900 rpm) at 37°C for 1 hour each. Restriction enzyme digests were then performed using DpnII. After enzymatic deactivation at 65°C, the resulting solution was diluted in 7 ml of water, and proximity ligation was performed using T4 ligase overnight. This was followed by overnight de-cross-linking using proteinase K. A second restriction enzyme digest was performed with Csp6i followed by a second proximity ligation step performed in 14 ml of solution. The resulting circularized library was extracted with ethanol and then purified using a HiPure polymerase chain reaction (PCR) cleanup kit. The cleaned library was then amplified using primers specific to *HP6/Umbrea* with attached Illumina P5/P7 adapters and sequenced on the Illumina HiSeq 2500 platform (PRJNA948431). Results were subsequently aligned to the FlyBase dmel6 reference genome, and raw coverage was visualized in IGV.

Fly stocks, genetic manipulations, and microscopy

All *D. melanogaster* lines were grown on a modified Bloomington cornmeal-molasses formulation. Fly lines for site-specific integration were obtained from Bloomington Drosophila Stock Center. pGreenRabbit reporter plasmids were site-specifically integrated into y[1] w[*] P{y[+t7.7] = nanos-phiC31\int.NLS}X; P{y[+t7.7] = CaryP}attP40 (BDSC 79604). FLEE1 (2 L:4444468-4450632) was amplified by PCR and cloned into the pGreenRabbit vector, following traditional cloning methods. We injected an empty pGreenRabbit vector as a negative control and pGreenRabbit with the *FLEE1* insert into BDSC 79604 pre-blastoderm embryos. Flies with successful integration were screened for the red eyes phenotype (presence of mini-white). We dissected salivary glands from third instar larvae of homozygous transformants in 1× phosphate-buffered saline (PBS), fixed in 5% paraformaldehyde in 1× PBS for 5 min, and washed four times in 1× PBS for 5 min. Fixed salivary glands were stained with 4′,6-diamidino-2-phenylindole (1:1000) for 10 min. All imaging was carried out on an upright laser scanning confocal microscope (Zeiss LSM 710) and similarly processed using ImageJ software.

Population genetic analysis

The data analysis

The genomic variants were called from whole-genome sequencing of 25 samples of *D. melanogaster* (DRM36, EA87, EA87N, ED10N, EF10N, EF126N, GA01, GA03, GA06, GA07, GH01, GH06, GH12, GH16, GH17, MC23, MC28, RAL900, RG18N, RG4N, UM118, UM37, UM526, ZH16, and ZH20), 10 samples of *D. simulans* (F11R4, F11R5, F21R2, F21R3, F31R2, F31R3, F31R4, F31R5, F41R1, and F41R2), and 5 samples of *D. yakuba* (CY02B5, CY08A, CY13A, CY17C, and CY22B), with sequencing depths >10× (*85*). All these publicly available raw reads were downloaded from National Center for Biotechnology Information and cleaned with fastp.0.23.4 (*86*). The cleaned reads were then mapped to the reference genome of BDGP6.32 with bwa mem v0.7.12 (*87*). The variants-calling was based on the practice of GATK4, including steps of marking duplicates, recalibrating base quality scores, per-sample calling with HaplotypeCaller, and joint-calling with GenotypeGVCFs. Single-nucleotide polymorphism (SNPs) annotation was performed with snpEff v5 (*85*, *88*). Only the biallelic sites with quality score >30, minimum coverage of 10×, minimum genotype quality of 30, and a maximum of 25% missing data were kept.

HKA-like tests (*30*, *89*) and MK tests (*31*) were conducted using polarized SNPs by focusing on fixed homologous sites in all outgroup samples (*D. yakuba* and *D. simulans*). The allele frequencies for *D. melanogaster* and outgroups were estimated with PLINK v1.9 (*90*). The expected proportions of diverged and polymorphic sites were calculated using the entirety of chromosome 2L (547951/307551 = 1.78). The proportions of diverged and polymorphic sites for genes were compared against the chromosome-wide ones with $\chi 2$ test (df = 1).

To detect signals of natural selection based on Ka/Ks (also ω) at the loci of *MFS18*, we collected orthologous sequences of these two genes in 10 *Drosophilid* species (*D. ananassae*, *Drosophila erecta*, *D. melanogaster*, *Drosophila mojavensis*, *D. pseudoobscura*, *D. simulans*, *Drosophila virilis*, *Drosophila willistoni*, *D. yakuba*, and *S. lebanonensis*) from OrthoDB v11 (*91*). For *HP6/Umbrea*, Ka/Ks ratio was not computed because of incomplete open reading frames in outgroup species. We used a codon-based alignment computed with TranslatorX and MAFFT v7.5 (*92*, *93*) for *MFS18* to generate gene trees and conducted the branch model test implemented by PAML v4.8 (*94*). To determine the optimal branch model for substitution rate estimation, we used a dynamic programming method by Zhang *et al.* (*29*) to select the optimal model according to log likelihoods.

The sojourn time of a neutral polymorphic duplicate before loss in a population

The question to address is how long a newly formed duplicate, if slightly deleterious [as was previously shown for various polymorphic duplicates (*1*)], can stay in a form of polymorphism in a population before loss due to genetic drift. The fixation probabilities for various polymorphic duplicates were calculated using the equation: $u/u_o = S/(1\text{-}e^{-S})$, where $u_o = 1/2Ne$ as the fixation probability of a neutral mutation, $S = 4N_es$ and s the selection coefficient (*95*). The

selection component, $\gamma = 2N_e s$, for various polymorphic duplicates in *D. melanogaster* were experimentally measured (*1*). The average sojourn time before a neutral duplicate mutation disappears from a population was calculated as $T_0(1/2\ N) = 2(Ne/N)\ln(2\ N)$, where Ne is effective population size and N actual population size (*96*). The average ratio Ne/N was reported in *D. melanogaster* as 0.027 (*97*) and a general estimate for metazoans as 0.10 (*98*). The $T_0(1/2\ N) <$ 1.04 ~ 3.60 generations (the median as 2.32 generations) [Ne = 3,300,000 in *D. melanogaster* (*2*)], because all the duplicate variants are slightly deleterious (*1*) and could disappear even sooner. Furthermore, the point mutation rate, as reported previously [e.g. (*2*, *4*)], is in the orders of 10^{-8} ~ 10^{-9} per site per generation, and the advantageous ones even much more rare, is unlikely to generate any genetic change that can rescue the duplicate from extinction in so short a time.

scRNA-seq

Testes from *D. melanogaster* (*42*), *D. yakuba* (newly generated), and *D. ananassae* (newly generated) were dissected in drops of cold PBS using forceps on petri dishes before being transferred on ice to reduce degradation. We then desheathed testes in lysis buffer [196 μl of 1× TrypLE + 4 μl of collagenase (100 mg/ml)]. After spinning down briefly and incubating at room temperature for 30 min with mild vortexing every 10 min, the samples were passed through 35-μm filters before centrifuging for 7 min at 163*g* (1200 rpm) at 4°C. We removed the supernatant, washed the cell pellet with 200 μl of cold Hanks' balanced salt solution (HBSS), and centrifuged again for 7 min at 163*g* (1200 rpm) at 4°C. We then removed the supernatant before resuspending the cell pellet in 35 μl of cold HBSS. We counted cells and checked viability on an automated cell counter using 5 μl of the single-cell suspension with 5 μl of trypan blue. The samples were then sent to Rockefeller Genomics Center for 10X single-cell library preparation and sequencing.

The resulting libraries were processed using cellranger (v7.1.0) and aligned to FlyBase genomes (dmel-r6.44, dyak-r1.05, dana-r1.06). Raw cell counts were loaded into Seurat (v. 4.1.3), and genomes were "melanogasterized" using one-to-one orthologs based on orthology information obtained from FlyBase ("dmel_orthologs_in_drosophila_species_fb_2022_01"). Cell identity information and 198-gene lists were obtained from (*43*), and new count matrices were generated using the 198-gene list as well as *HP6/Umbrea* and coexpression cluster genes (excluding CG42523 due to QC issues). As the *mel-ana* divergence predated *HP6/Umbrea* origination, *D. ananassae* counts for *HP6/Umbrea* were generated and set uniformly to 0. Subsequent data were processed using these count matrices, while dimensional reduction and projection (principal components analysis + Uniform Manifold Approximation and Projection) was performed solely on the 198-gene list. Presented data, including distributions and violin plots, used raw count data.

RNAi and lethality measurements

We used lethality data previously published by our laboratory (*22*, *23*) that were based on RNA interference (RNAi) lines obtained from the Vienna Drosophila Resource Center (VDRC). A quarter of all KK RNAi lines from VDRC carry an inverted repeat sequence insertion at 30B3. However, a proportion (23 to 25%) of KK lines also carry an insertion at 40D3, which is housed within the *tio* locus and produces a confounding lethal phenotype. To avoid this, we updated the lethality data of new genes reported in (*23*) by removing the *tio* insertion site in KK lines using a recombination-based approach (*22*, *99*) and finally derived lethality data for the new genes. The lethality results for all lines without insertion in the *tio* locus were reproducible, previously having been analyzed using four replicates, and again in our analysis in duplicate. Distally duplicated genes had 90% fewer offspring relative to control flies after Act5c-GAL4 induction were labeled as essential.

Supplementary Materials

The PDF file includes:

Supplementary Text
Figs. S1 to S8
Legends for tables S1 and S2
Legend for movie S1
Data S1

Other Supplementary Material for this manuscript includes the following:

Tables S1 and S2
Movie S1

REFERENCES AND NOTES

1. J. J. Emerson, M. Cardoso-Moreira, J. O. Borevitz, M. Long, Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science* **320**, 1629–1631 (2008).
2. M. Kreitman, Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417 (1983).
3. U. Bergthorsson, D. I. Andersson, J. R. Roth, Ohno's dilemma: Evolution of new genes under continuous selection. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17004–17009 (2007).
4. Y. Wang, P. McNeil, R. Abdulazeez, M. Pascual, S. E. Johnston, P. D. Keightley, D. J. Obbard, Variation in mutation, recombination, and transposition rates in *Drosophila melanogaster* and *Drosophila simulans*. *Genome Res.* **33**, 587–598 (2023).
5. A. Force, M. F. Lynch, B. Pickett, A. Amores, Y. Yan, J. Postlethwait, Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
6. C. T. Hittinger, S. B. Carroll, Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**, 677–681 (2007).
7. J. Nasvall, L. Sun, J. R. Roth, D. I. Andersson, Real-time evolution of new genes by innovation, amplification, and divergence. *Science* **338**, 384–387 (2012).
8. Y. Eguchi, K. Makanae, T. Hasunuma, Y. Ishibashi, K. Kito, H. Moriya, Estimating the protein burden limit of yeast cells by measuring the expression limits of glycolytic proteins. *eLife* **7**, e34595 (2018).
9. A. M. Rice, A. McLysaght, Dosage-sensitive genes in evolution and disease. *BMC Biol.* **15**, 78 (2017).
10. Q. Zhou, G. Zhang, Y. Zhang, S. Xu, R. Zhao, Z. Zhan, X. Li, Y. Ding, S. Yang, W. Wang, On the origin of new genes in Drosophila. *Genome Res.* **18**, 1446–1455 (2008).
11. A. Larkin, S. J. Marygold, G. Antonazzo, H. Attrill, G. D. Santos, P. V. Garapati, J. L. Goodman, L. S. Gramates, G. Millburn, V. B. Strelets, C. J. Tabone, J. Thurmond, FlyBase Consortium, FlyBase: Updates to the *Drosophila melanogaster* knowledge base. *Nucleic Acids Res.* **49**, D899–D907 (2021).
12. B. R. Graveley, A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin, L. Yang, C. G. Artieri, M. J. van Baren, N. Boley, B. W. Booth, J. B. Brown, L. Cherbas, C. A. Davis, A. Dobin, R. Li, W. Lin, J. H. Malone, N. R. Mattiuzzo, D. Miller, D. Sturgill, B. B. Tuch, C. Zaleski, D. Zhang, M. Blanchette, S. Dudoit, B. Eads, R. E. Green, A. Hammonds, L. Jiang, P. Kapranov, L. Langton, N. Perrimon, J. E. Sandler, K. H. Wan, A. Willingham, Y. Zhang, Y. Zou, J. Andrews, P. J. Bickel, S. E. Brenner, M. R. Brent, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, B. Oliver, S. E. Celniker, The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**, 473–479 (2011).
13. H. Kaessmann, N. Vinckenbosch, M. Long, RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat. Rev. Genet.* **10**, 19–31 (2009).
14. D. Jangam, C. Feschotte, E. Betran, Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.* **33**, 817–831 (2017).
15. J. B. Brown, N. Boley, R. Eisman, G. E. May, M. H. Stoiber, M. O. Duff, B. W. Booth, J. Wen, S. Park, A. M. Suzuki, K. H. Wan, C. Yu, D. Zhang, J. W. Carlson, L. Cherbas, B. D. Eads, D. Miller, K. Mockaitis, J. Roberts, C. A. Davis, E. Frise, A. S. Hammonds, S. Olson, S. Shenker, D. Sturgill, A. A. Samsonova, R. Weiszmann, G. Robinson, J. Hernandez, J. Andrews, P. J. Bickel, P. Carninci, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, E. C. Lai, B. Oliver, N. Perrimon, B. R. Graveley, S. E. Celniker, Diversity and dynamics of the *Drosophila* transcriptome. *Nature* **512**, 393–399 (2014).

16. T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, G. Cavalli, Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
17. G. Fasano, A. Franceschini, A multidimensional version of the Kolmogorov–Smirnov test. *Mon. Not. R. Astron. Soc.* **225**, 155–170 (1987).
18. C. Puritz, E. Ness-Cohn, R., Braun, fasano.franceschini.test: An implementation of a multivariate KS test in R. *R J.* **15**, 159–171 (2023).
19. B. D. Ross, L. Rosin, A. W. Thomae, M. A. Hiatt, D. Vermaak, A. F. A. de la Cruz, A. Imhof, B. G. Mellone, H. S. Malik, Stepwise evolution of essential centromere function in a *Drosophila* neogene. *Science* **340**, 1211–1214 (2013).
20. S. Grill, A. Riley, M. Selvaraj, R. Lehmann, HP6/Umbrea is dispensable for viability and fertility, suggesting essentiality of newly evolved genes is rare. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2309478120 (2023).
21. F. Greil, E. de Wit, H. J. Bussemaker, B. van Steensel, HP1 controls genomic targeting of four novel heterochromatin proteins in *Drosophila*. *EMBO J.* **26**, 741–751 (2007).
22. S. Xia, N. W. VanKuren, C. Chen, L. Zhang, C. Kemkemer, Y. Shao, H. Jia, U. Lee, A. S. Advani, A. Gschwent, M. D. Vibranovski, S. Chen, Y. E. Zhang, M. Long, Genomic analyses of new genes and their phenotypic effects reveal rapid evolution of essential functions in *Drosophila* development. *PLOS Genet.* **17**, e1009654 (2021).
23. S. Chen, Y. E. Zhang, M. Long, New genes in *Drosophila* quickly become essential. *Science* **330**, 1682–1685 (2010).
24. S. E. Celniker, L. A. L. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. MacAlpine, G. Micklem, F. Piano, M. Snyder, L. Stein, K. P. White, R. H. Waterston, modENCODE Consortium, Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
25. M. I. Krzywinski, J. E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M. A. Marra, Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
26. S. Heinz, L. Taxari, M. G. B. Hayes, M. Urbanowski, M. W. Chang, N. Givarkes, A. Rialdi, K. M. White, R. A. Albrecht, L. Pache, I. Marazzi, A. Garcia-Sastre, M. L. Shaw, C. Benner, Transcription elongation can affect genome 3D structure. *Cell* **174**, 1522–1536.e22 (2018).
27. B. E. Housden, K. Millen, S. J. Bray, *Drosophila* reporter vectors compatible with ΦC31 integrase transgenesis techniques and their use to generate new notch reporter fly lines. *G3* **2**, 79–82 (2012).
28. C. Joppich, S. Scholz, G. Korge, A. Schwendemann, Umbrea, a chromo shadow domain protein in *Drosophila melanogaster* heterochromatin, interacts with Hip, HP1 and HOAP. *Chromosome Res.* **17**, 19–36 (2009).
29. C. Zhang, J. Wang, W. Xie, G. Zhou, M. Long, Q. Zhang, Dynamic programming procedure for searching optimal models to estimate substitution rates based on the maximum-likelihood method. *Proc. Natl. Acad. Sci.* **108**, 7860–7865 (2011).
30. R. R. Hudson, M. Kreitman, M. Aguadé, A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159 (1987).
31. J. H. McDonald, M. Kreitman, Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
32. M. Z. Ludwig, C. Bergman, N. H. Patel, M. Kreitman, Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564–567 (2000).
33. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
34. M. T. Weirauch, A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J.-C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. M. Walhout, F.-Y. Bouget, G. Ratsch, L. F. Larrondo, J. R. Ecker, T. R. Hughes, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
35. S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. R. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White, M. Kellis, Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
36. L. S. Gramates, J. Agapite, H. Attrill, B. R. Calvi, M. A. Crosby, G. dos Santos, J. L. Goodman, D. Goutte-Gattat, V. K. Jenkins, T. Kaufman, A. Larkin, B. B. Matthews, G. Millburn, V. B. Strelets, the FlyBase Consortium, FlyBase: A guided tour of highlighted features. *Genetics* **220**, iyac035 (2022).
37. M. Acar, H. Jafar-Nejad, N. Giagtzoglou, S. Yallampalli, G. David, Y. He, C. Delidakis, H. J. Bellen, Senseless physically interacts with proneural proteins and functions as a transcriptional co-activator. *Development* **133**, 1979–1989 (2006).
38. M. Long, C. H. Langley, Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science* **260**, 91–95 (1993).
39. M. D. Vibranovski, Y. E. Zhang, C. Kemkemer, H. F. Lopes, T. L. Karr, M. Long, Re-analysis of the larval testis data on meiotic sex chromosome inactivation revealed evidence for tissue-specific gene expression related to the drosophila X chromosome. *BMC Biol.* **10**, 49 (2012).
40. W. Zhang, P. Landback, A. R. Gschwend, B. Shen, M. Long, New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol.* **16**, 202 (2015).
41. M. Long, N. W. VanKuren, S. Chen, M. D. Vibranovski, New gene evolution: Little did we know. *Annu. Rev. Genet.* **47**, 307–333 (2013).
42. E. Witt, S. Benjamin, N. Svetec, L. Zhao, Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in *Drosophila*. *eLife* **8**, e47138 (2019).
43. U. Lee, C. Li, C. B. Langer, N. Svetec, L. Zhao, Comparative single cell analysis of transcriptional bursting reveals the role of genome organization on de novo transcript origination. bioRxiv 591771 [Preprint] (2024). https://doi.org/10.1101/2024.04.29.591771.
44. R. Y. Birnbaum, E. J. Clowney, O. Agamy, M. J. Kim, J. Zhao, T. Yamanaka, Z. Pappalardo, S. L. Clarke, A. M. Wenger, L. Nguyen, F. Gurrieri, D. B. Everman, C. E. Schwartz, O. S. Birk, G. Bejerano, S. Lomvardas, N. Ahituv, Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res.* **22**, 1059–1068 (2012).
45. S. Zhenilo, E. Khrameeva, S. Tsygankova, N. Zhigalova, A. Mazur, E. Prokhortchouk, Individual genome sequencing identified a novel enhancer element in exon 7 of the CSFR1 gene by shift of expressed allele ratios. *Gene* **566**, 223–228 (2015).
46. P. T. Spellman, G. M. Rubin, Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* **1**, 5 (2002).
47. C. P. Johnstone, K. E. Galloway, Supercoiling-mediated feedback rapidly couples and tunes transcription. *Cell Rep.* **41**, 111492 (2022).
48. T. Fukaya, B. Lim, M. Levine, Enhancer control of transcriptional bursting. *Cell* **166**, 358–368 (2016).
49. R. Loganathan, J. H. Kim, M. B. Wells, D. J. Andrew, Secrets of secretion-How studies of the *Drosophila* salivary gland have informed our understanding of the cellular networks underlying secretory organ form and function. *Curr. Top. Dev. Biol.* **143**, 1–36 (2021).
50. D. M. Johnson, M. B. Wells, R. Fox, J. S. Lee, R. Loganathan, D. Levings, A. Bastien, M. Slattery, D. J. Andrew, CrebA increases secretory capacity through direct transcriptional regulation of the secretory machinery, a subset of secretory cargo, and other key regulators. *Traffic* **21**, 560–577 (2020).
51. R. M. Fox, C. D. Hanlon, D. J. Andrew, The CrebA/Creb3-like transcription factors are major and direct regulators of secretory capacity. *J. Cell Biol.* **191**, 479–492 (2010).
52. R. M. Fox, A. Vaishnavi, R. Maruyama, D. J. Andrew, Organ-specific gene expression: The bHLH protein Sage provides tissue specificity to *Drosophila* FoxA. *Development* **140**, 2160–2171 (2013).
53. E. W. Abrams, M. S. Vining, D. J. Andrew, Constructing an organ: The *Drosophila* salivary gland as a model for tube formation. *Trends Cell Biol.* **13**, 247–254 (2003).
54. R. Dittrich, T. Bossing, A. P. Gould, G. M. Technau, J. Urban, The differentiation of the serotonergic neurons in the *Drosophila* ventral nerve cord depends on the combined function of the zinc finger proteins Eagle and Huckebein. *Development* **124**, 2515–2525 (1997).
55. W. Wang, J. Zhang, C. Alvarez, A. Llopart, M. Long, The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in *Drosophila melanogaster*. *Mol. Biol. Evol.* **17**, 1294–1301 (2000).
56. W. Wang, H. Zheng, S. Yang, Y. Haijing, J. Li, H. Jiang, J. Su, L. Yang, J. Zhang, J. McDermott, R. Samudrala, J. Wang, H. Yang, J. Yun, K. Kristiansen, G. K. S. Wong, J. Wang, Origin and evolution of new exons in rodents. *Genome Res.* **15**, 1258–1264 (2005).
57. C. B. Bridges, The bar "gene" a duplication. *Science* **83**, 210–211 (1936).
58. H. J. Muller, Bar duplication. *Science* **83**, 528–530 (1936).
59. N. Harmston, E. Ing-Simmons, G. Tan, M. Perry, M. Markenschlager, B. Lenhard, Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat. Commun.* **8**, 441 (2017).
60. J. Krefting, M. A. Andrade-Navarro, J. Ibn-Salen, Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biol.* **16**, 87 (2018).
61. J. Y. Zhang, Q. Zhou, On the regulatory evolution of new genes throughout their life history. *Mol. Biol. Evol.* **36**, 15–27 (2019).
62. H. Dai, T. F. Yoshimatsu, M. Long, Retrogene movement within- and between-chromosomes in the evolution of *Drosophila* genomes. *Gene* **385**, 96–102 (2006).

63. L. E. Kursel, H. McConnel, A. F. A. de la Cruz, H. S. Malik, Gametic specialization of centromeric histone paralogs in *Drosophila virilis*. *Life Sci. Alliance* **4**, e202000992 (2021).
64. N. W. VanKuren, M. Long, Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nat. Ecol. Evol.* **2**, 705–712 (2018).
65. L. Zech, U. Haglund, K. Nilsson, G. Klein, Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with burkitt and non-burkitt lymphomas. *Int. J. Cancer* **17**, 47–56 (1976).
66. V. K. Jain, J. G. Judde, E. E. Max, I. T. Magrath, Variable IgH chain enhancer activity in Burkitt's lymphomas suggests an additional, direct mechanism of c-myc deregulation. *J. Immunol.* **150**, 5418–5428 (1993).
67. B. Gryder, P. C. Scacheri, T. Ried, J. Khan, Chromatin mechanisms driving cancer. *Cold Spring Harb. Perspect. Biol.* **14**, a040956 (2022).
68. J. Zheng, Oncogenic chromosomal translocations and human cancer (review). *Oncol. Rep.* **30**, 2011–2019 (2013).
69. R. C. Hennessey, K. M. Brown, Cancer regulatory variation. *Curr. Opin. Genet. Dev.* **66**, 41–49 (2021).
70. X. Wang, J. Xu, B. Zhang, Y. Hou, F. Song, H. Lyu, F. Yue, Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat. Methods* **18**, 661–668 (2021).
71. L. E. Montefiori, S. Bendig, Z. Gu, X. Chen, P. Pölönen, X. Ma, A. Murison, A. Zeng, L. Garcia-Prat, K. Dickerson, I. Iacobucci, S. Abdelhamed, R. Hiltenbrand, P. E. Mead, C. M. Mehr, B. Xu, Z. Cheng, T.-C. Chang, T. Westover, J. Ma, A. Stengel, S. Kimura, C. Qu, M. B. Valentine, M. Rashkovan, S. Luger, M. R. Litzow, J. M. Rowe, M. L. den Boer, V. Wang, J. Yin, S. M. Kornblau, S. P. Hunger, M. L. Loh, C.-H. Pui, W. Yang, K. R. Crews, K. G. Roberts, J. J. Yang, M. V. Relling, W. E. Evans, W. Stock, E. M. Paietta, A. A. Ferrando, J. Zhang, W. Kern, T. Haferlach, G. Wu, J. E. Dick, J. M. Klco, C. Haferlach, C. G. Mullighan, Enhancer hijacking drives oncogenic BCL11B expression in lineage-ambiguous stem cell leukemia. *Cancer Discov.* **11**, 2846–2867 (2021).
72. S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X.-W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, A. M. Chinnaiyan, Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
73. M. S. Litwin, H.-J. Tan, The diagnosis and treatment of prostate cancer: A review. *JAMA* **317**, 2532–2542 (2017).
74. K. J. Meaburn, T. Misteli, E. Soutoglou, Spatial genome organization in the formation of chromosomal translocations. *Semin. Cancer Biol.* **17**, 80–90 (2007).
75. F. Dubois, N. Sidiropoulos, J. Weischenfeldt, R. Beroukhim, Structural variations in cancer and the 3D genome. *Nat. Rev. Cancer* **22**, 533–546 (2022).
76. H. Neves, C. Ramos, M. G. da Silva, A. Parreira, L. Parreira, The nuclear topography of ABL, BCR, PML, and RARalpha genes: Evidence for gene proximity in specific phases of the cell cycle and stages of hematopoietic differentiation. *Blood* **93**, 1197–1207 (1999).
77. S. Kozubek, E. Lukásová, A. Marecková, M. Skalníková, M. Kozubek, E. Bártová, V. Kroha, E. Krahulcová, J. Slotová, The topological organization of chromosomes 9 and 22 in cell nuclei has a determinative role in the induction of t(9,22) translocations and in the pathogenesis of t(9,22) leukemias. *Chromosoma* **108**, 426–435 (1999).
78. C. S. Osborne, L. Chakalova, J. A. Mitchell, A. Horton, A. L. Wood, D. J. Bolland, A. E. Corcoran, P. Fraser, Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. *PLOS Biol.* **5**, e192 (2007).
79. S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, C. K. Glass, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
80. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
81. C. Keil, R. W. Leach, S. M. Faizaan, S. Bezawada, L. Parsons, A. Baryshnikova, Treeview 3.0 (beta 1) - Visualization and analysis of large data matrices, Zenodo (2018); https://doi.org/10.5281/zenodo.1303402.
82. A. Shumate, S. L. Salzberg, Liftoff: Accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
83. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
84. A. Shumate, S. Salzberg, LiftoffTools: A toolkit for comparing gene annotations mapped between genome assemblies. *F1000Res.* **11**, 1230 (2022).
85. J. M. Coughlan, A. J. Dagilis, A. Serrato-Capuchina, H. Elias, D. Peede, K. Isbell, D. M. Castillo, B. S. Cooper, D. R. Matute, Patterns of population structure and introgression among recently differentiated Drosophila melanogaster populations. *Mol. Biol. Evol.* **39**, msac223 (2022).
86. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
87. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
88. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
89. D. J. Begun, A. K. Holloway, K. Stevens, L. W. Hillier, Y.-P. Poh, M. W. Hahn, P. M. Nista, C. D. Jones, A. D. Kern, C. N. Dewey, L. Pachter, E. Myers, C. H. Langley, Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila* simulans. *PLOS Biol.* **5**, e310 (2007).
90. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
91. D. Kuznetsov, F. Tegenfeldt, M. Manni, M. Seppey, M. Berkeley, E. V. Kriventseva, E. M. Zdobnov, OrthoDB v11: Annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**, D445–D451 (2023).
92. F. Abascal, R. Zardoya, M. J. Telford, TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–W13 (2010).
93. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
94. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
95. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 2012); https://cambridge.org/core/books/neutral-theory-of-molecular-evolution/0FF60E9F47915B17FFA2620C49400632.
96. D. L. Hartl, A. G. Clark, *Principles of Population Genetics* (Oxford Univ. Press, ed. 2, 2006).
97. D. A. Briscoe, J. M. Malpica, A. Robertson, G. J. Smith, R. Frankham, R. G. Banks, J. S. F. Barker, Rapid loss of genetic variation in large captive populations of Drosophila flies: Implications for genetic management of captive populations. *Conservation Biology* **6**, 416–425 (1992).
98. R. Frankham, Effective population size/adult population size ratios in wildlife: A review. *Genet. Res.* **66**, 95–107 (1995).
99. E. W. Green, G. Fedele, F. Giorgini, C. P. Kyriacou, A *Drosophila* RNAi collection is subject to dominant phenotypic effects. *Nat. Methods* **11**, 222–223 (2014).

Acknowledgments: We thank M. Halfon and S. Bray for sharing the pGreenRabbit plasmid. We thank H. Duan and C. Zhao at the Genomics Resource Center of Rockefeller University for help with the scRNA-seq libraries. Figures 1 and 3 to 7 were partially designed using BioRender. This manuscript is dedicated in memory of F. Lee. **Funding:** This work was supported by the National Institutes of Health grant GM7197-42 (U.L.), National Science Foundation postdoctoral fellowship 2410289 (U.L.), National Institutes of Health grant 1R01GM116113-01A1 (M.L.), National Science Foundation MCB2020667 (M.L.), National Institutes of Health postdoctoral fellowship F32GM146423 (D.A.), National Institutes of Health grant R01-GM115523 (P.A.), and National Institutes of Health MIRA R35GM133780 (L. Zhao). **Author contributions:** Conceptualization: U.L. and M.L. Methodology: U.L. and M.L. Investigation: U.L., D.A., S.X., M.A., D.R.S., I.E., D.S., J.C., P.R., N.S., C.L., C.B.L., J.J.E., and L. Zhang. Visualization: U.L., D.A., and P.R. Funding acquisition: P.A., Q.Z., L. Zhao, and M.L. Project administration: U.L., P.A., Q.Z., L. Zhao, and M.L. Supervision: P.A., Q.Z., J.J.E., L. Zhao, and M.L. Writing—original draft: U.L., D.A., A.G., and M.L. Writing—review and editing: U.L., D.A., P.R., A.G., N.S., L. Zhao, and M.L. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** *D. melanogaster* Hi-C libraries form S2 cells: PRJNA393992 are publicly available. *D. teissieri* Hi-C libraries from adult females: SRR12331760 are publicly available. *D. pseudoobscura* Hi-C library from L3 larvae: PRJNA948678 is newly generated. *D. pseudoobscura* reference genome: PRJNA596268 is publicly available. *D. melanogaster* 4C-Seq data from L3 larve: PRJNA948431 are newly generated. scRNA-seq data: PRJNA995212. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Scripts for analyses may be found at https://github.com/ulee-sciscripts/enh_cap_div_scripts and https://doi.org/10.5061/dryad.x69p8czv0.

Submitted 20 December 2023
Accepted 11 November 2024
Published 18 December 2024
10.1126/sciadv.adn6625