

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

UnJin Lee *et al.*

Corresponding author: UnJin Lee, [ulee@rockefeller.edu](mailto:ulee@rockefeller.edu); Manyuan Long, [mlong@uchicago.edu](mailto:mlong@uchicago.edu)

*Sci. Adv.* **10**, eadn6625 (2024)  
DOI: 10.1126/sciadv.adn6625

**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

## SUPPLEMENTARY INFORMATION:

### Prior Models and Genomic Symmetries

The first models describing new gene evolution proposed that all new genes likely evolve via duplication-based mechanisms (58, 99), including: the duplication, divergence, complementation (DDC)/sub-functionalization model (5), the escape from adaptive conflict (EAC) model (6), the innovation, amplification, and divergence (IAD) model. To address how a duplicate, redundant gene copy may rise to fixation, these models all assume multiple functions for any studied gene. In the DDC (sub-functionalization) model, symmetric (identical) copies of a duplicated gene lose function in complementary fashion, resulting in retention of duplicate gene copies with separate but complementary functions. While the DDC model allows each duplicate copy to possess a subset of the parental gene's original functions, the EAC model allows for increased optimization of one or more of the original parental gene's functions that are partitioned to each paralogous copy. The EAC model assumes internal genetic conflict within the parental gene preventing simultaneous optimization of its multiple functions, and duplication thus allows for the resolution of this evolutionary constraint, conferring a selective advantage in both parental and new genes. While the DDC and EAC models can explain how prior gene functions can be partitioned amongst duplicate copies, these models both assume that newly evolved duplicated genes can only retain pre-existing, essential functions from their parental genes and thus fail to describe a mechanism for how truly evolutionarily novel gene function emerges. In contrast, the IAD model proposes that changes in selection pressures may favor the increased expression of a given gene with an auxiliary function. This provides a selective advantage for increased gene dosage through an increase in gene copy number. Following the initial increase of auxiliary function through gene amplification, subsequent relaxation of selection pressure will allow for changes to accumulate on the various copies, allowing the new copies to diverge and potentially gain a new function (3). While the IAD model provides a solution for Ohno's dilemma for gene family expansions in microbial organisms while encountering environmental changes (7), the model cannot be applied to metazoans due to often conflict effects for same genes in different tissues or cells.

A key factor missing in these previous models is the effect of chromosomal and regulatory context on a gene duplicate's function and spatiotemporal expression. In the DDC, EAC and IAD models, the evolution of new gene duplicates is assumed to occur in a regulatory-independent context and do not describe how the regulatory sequences may shape the evolution of a new gene duplicate. Here, we explain how the regulatory context can promote neofunctionalization of newly duplicated genes through the enhancer-capture divergence (ECD) model. In the ECD model, the duplication of a pre-existing gene into a new regulatory context through a preexisting 3-dimensional (3D) genome structure results in unique expression pattern from that of its parent gene controlled by a combination of regulatory elements from both the native and new contexts. The single-step evolutionary process of ECD thus allows for rapid neofunctionalization and is dependent on the regulatory architecture of the three-dimensional eukaryotic genome.

Similar to the IAD model, the ECD model first proposes that selective pressures change for the increased expression of a pre-existing (parental) gene within a specific tissue or set of tissues. To achieve this, there are two possible scenarios: 1) the evolution of a new enhancer in the parental gene's locus, either through duplication or substitution, or in the case of the ECD model, 2) the duplication of the parental gene into a distal region of the genome that is already under the control of a pre-existing, tissue-specific enhancer. While the first scenario is possible, this would require

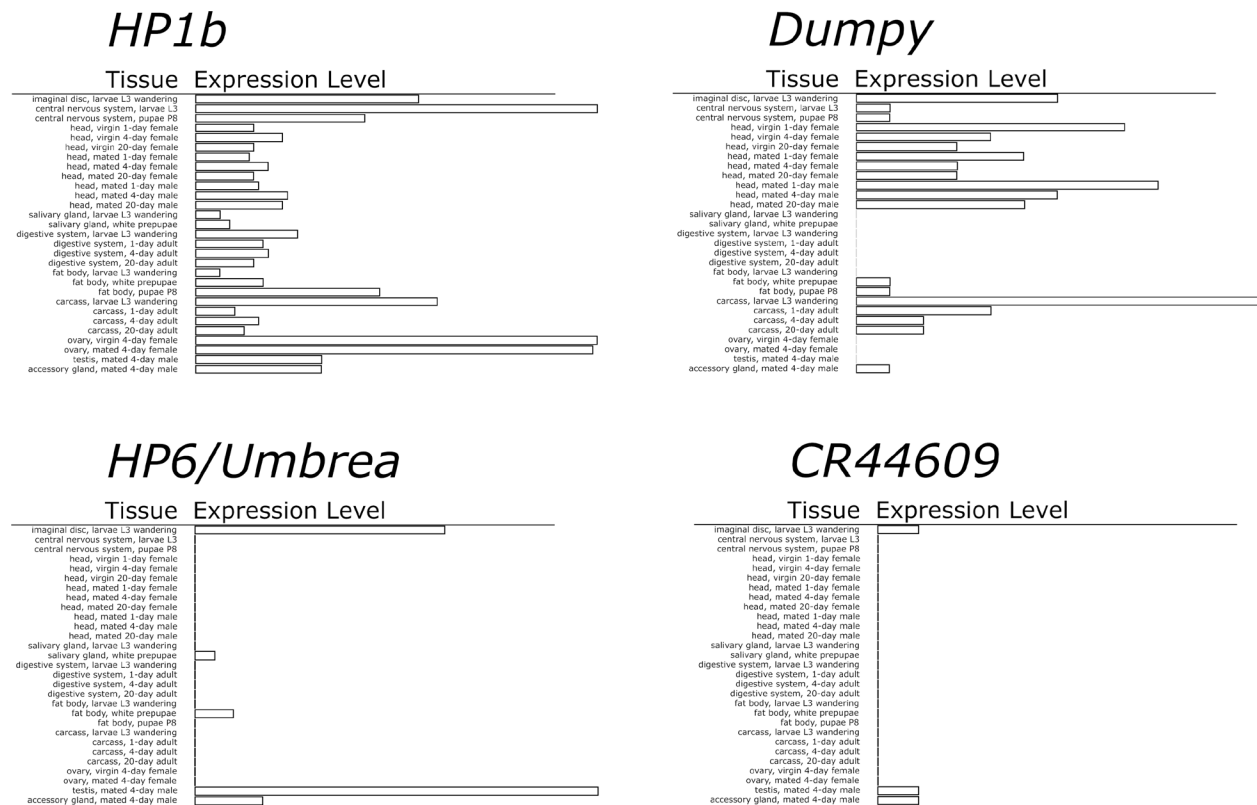
multiple neutral *de novo* substitutions or insertions to generate one or more necessary transcription factor binding sites that fix within a population and modulates the expression of the new gene duplicate without disrupting parent gene's expression pattern.

In the second scenario under *enhancer capture*, the duplication of the parental gene into another regulatory environment under the control of a pre-existing, 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 that is under new regulatory control will increase in frequency in the population, allowing it to fix. If the selection pressures change such that the increased tissue-specific expression of the new gene is no longer advantageous or 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 function.

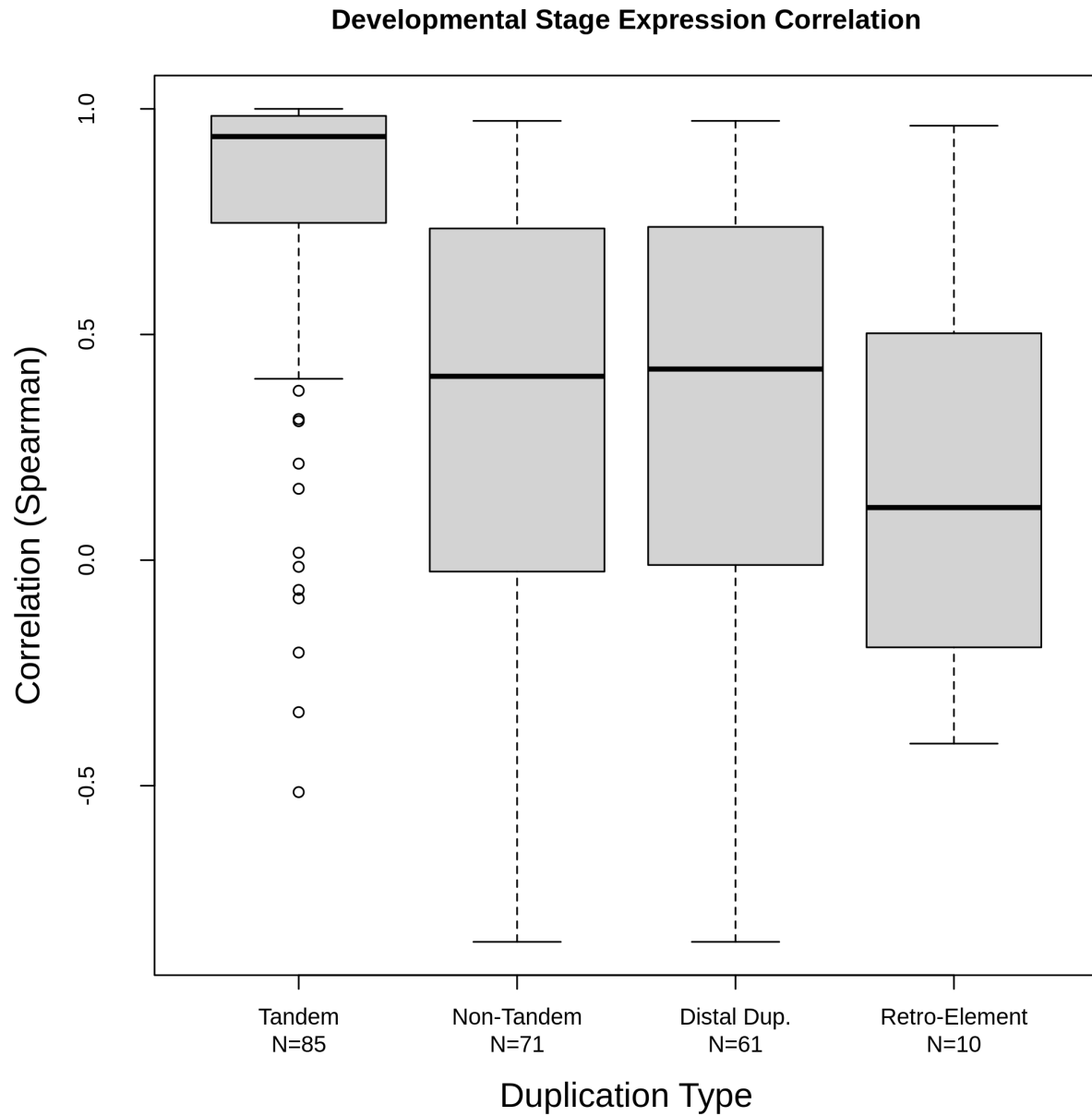
There are several distinctions between the ECD model and previously classic models of gene duplication. First, the DDC, EAC and IAD models do not consider the effect of the pre-existing regulatory and chromosomal environment on a new, distally duplicated gene. Second, compared to the DDC and EAC models but like IAD, the ECD model is a single-step process in which the initial duplication event provides a selective advantage. However, unlike IAD, a duplicate gene copy can immediately integrate into a tissue-specific regulatory network separate from that of its parent under ECD, providing a fast evolutionary solution to "Ohno's Dilemma." A final and critical distinction between the ECD model and previous classical models, to address the dilemma, is that they explain the evolution and retention of different classes of gene duplications. The previous models 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. A similar genomic symmetry is also seen in tandem duplications, where duplicate copies cannot be definitively identified as the "parent" or "new" gene copy through synteny. As a result, the DDC, EAC, and IAD models provide reasonable mechanistic explanations for why a large number of duplicate gene copies are retained, applying particularly well to tandem and other symmetric gene duplications. However, these previous models do not consider the role of regulatory and chromosomal context on newly evolved, asymmetric duplicates and thus cannot explain the origination of a large number of evolutionarily important genes.

Genes evolving under ECD are asymmetric, as the parental gene remains in its original locus while the new copy resides in a distal region of the genome under the control of a different, pre-existing regulatory context. This genomic asymmetry allows for clear distinction between parent and new gene copies through synteny. A similar asymmetry is also seen in protein and regulatory function, where the parent gene retains its entire function and spatiotemporal expression pattern, while the auxiliary tissue-specific function and expression pattern is restricted to the new gene copy. The asymmetry of both 1) distinguishable gene identity and 2) segregation of expression and function is a key feature of the ECD model that distinguishes it from the DDC, EAC and IAC models, and allows for clear identification of genes that evolved under enhancer capture and the application of genomic tests regarding retention of essential gene function. We utilize these features of the ECD model to show a statistical enrichment of distally duplicated genes that have evolved via enhancer capture-divergence within *Drosophila melanogaster*. Under the ECD model, we predict that newly evolved genes will be enriched for two elements: 1) we predict high degrees of co-expression with neighboring genes combined with low co-expression with its parent gene

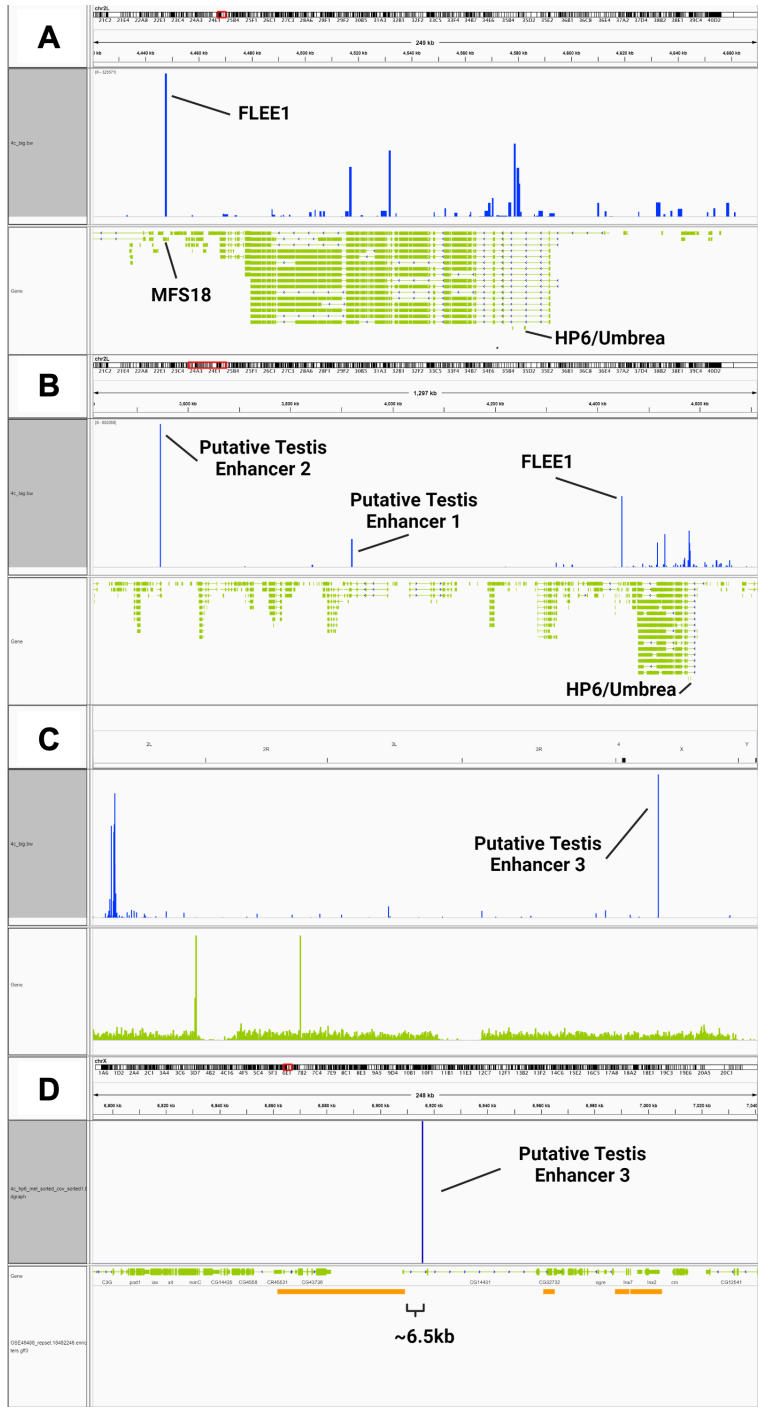
and 2) we predict that genes evolving under ECD should originate as non-essential, as all essential function should be asymmetrically retained by the parent gene while the auxiliary, non-essential function is retained in the new copy. As the ECD process can occur in a single step in a 3D world of genome, we also predict that the enhancer capture process should be a key mechanism for the evolution of distally duplicated genes alongside the DDC, EAC, and IAD models due to its rapid evolvability. This is supported by the observation that gene duplication occurs more frequently than point mutation (3, 7), where ECD requires fewer genomic alterations than the *de novo* evolution of a new enhancer via substitution.



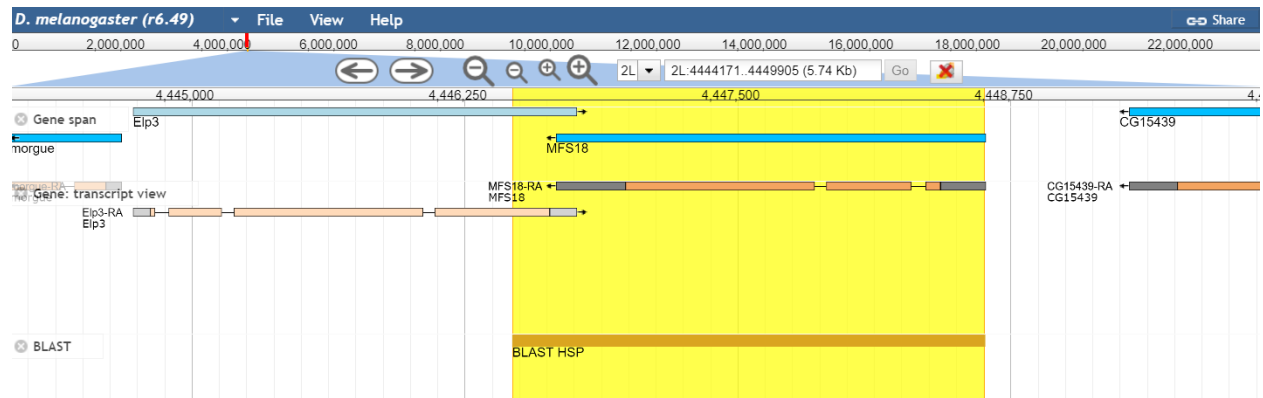
**Figure S1. Expression patterns of *HP6/Umbrea* and other genes.** Unlike the broad expression pattern of parental gene *HP1b*, the tissue expression pattern of *HP6/Umbrea* is stereotypical of new gene expression patterns, with high tissue specificity, restricted in this case to primarily the imaginal discs, larval salivary glands, and male reproductive organs. While *HP6/Umbrea* was inserted into an intronic region of the larger gene *dumpy*, *HP6/Umbrea*’s expression pattern is shared with *HP6/Umbrea*’s neighboring gene *CR44609*.



**Figure S2. Enhancer Capture-Divergence drives regulatory neo-functionalization of new duplicate genes.** Parental and new gene co-expression for new duplicate genes arising by tandem, distal duplicates, retro-transposons, and non-tandem (distal + retro-transposons) duplicates were calculated using gene expression data for 30 developmental stages in *D. melanogaster* (“developmental co-expression”). The development co-expression of non-tandem duplicates was significantly lower than the developmental co-expression of tandem duplicates ( $p=3.45 \times 10^{-10}$ ) as well as distal duplicates and retro-transposons alone (distal:  $p=8.99 \times 10^{-9}$ , retro-transposition:  $p=5.41 \times 10^{-3}$ ). These combined results demonstrate how Enhancer Capture-Divergence is a significant driver of regulatory neo-functionalization in new duplicate genes, which cannot be explained by symmetric models of new duplicate gene evolution.

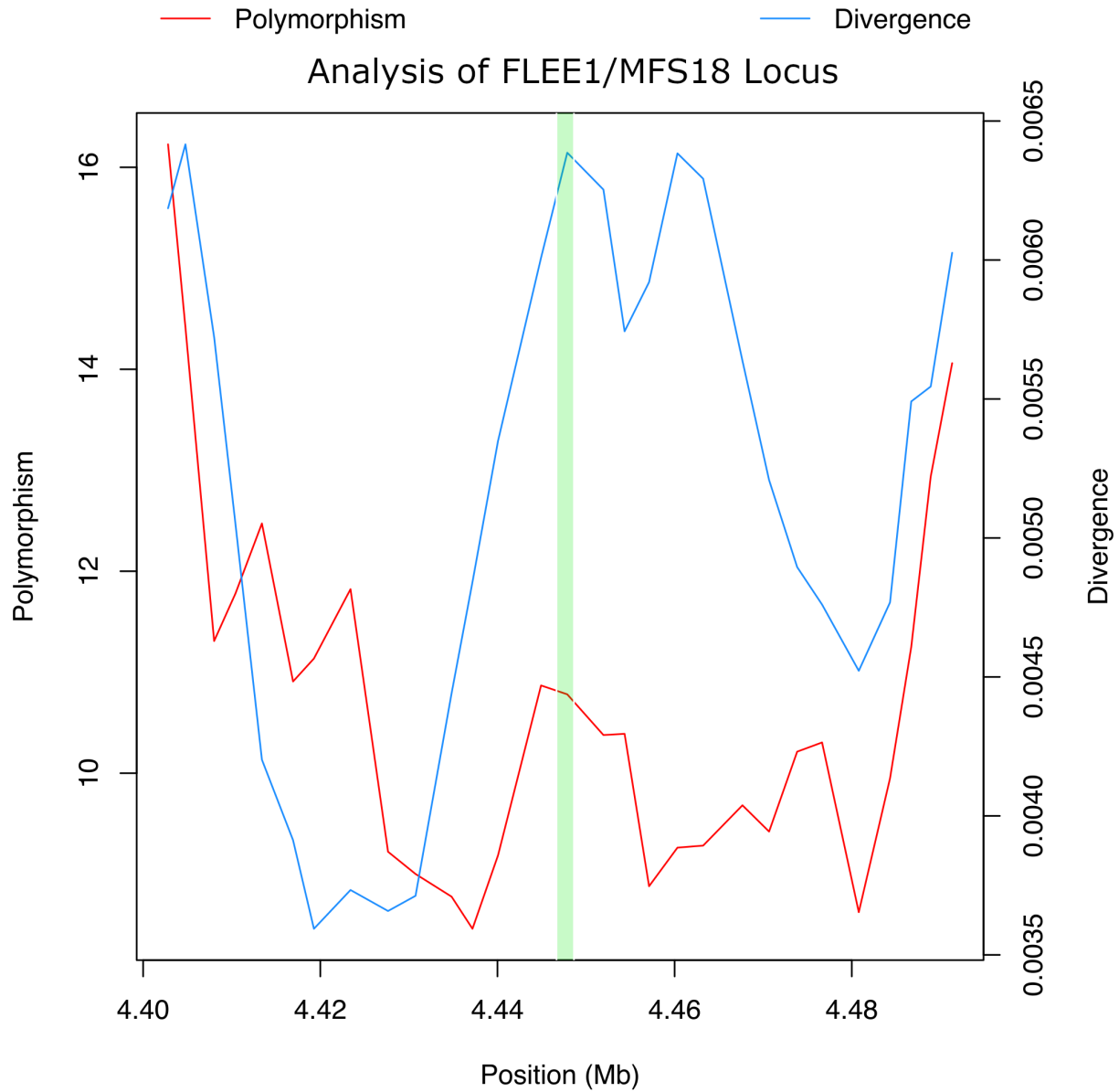


**Figure S3. 4C-Seq in *D. melanogaster*.** Shown above are raw read coverage results from 4C-Seq derived from *D. melanogaster* larval tissue as visualized in IGV. Self-self interactions have been removed. **(a)** The location of the tested FLEE1 enhancer is shown relative to *HP6/Umbrea* and *MFS18*. **(b)** The untested Putative Testis Enhancer 1 and 2 loci (Chr2L:3919608-3920575, Chr2L:3545704-3545771 respectively) are shown relative to *HP6/Umbrea*. **(c)** The untested inter-chromosomal Putative Testis Enhancer 3 locus (ChrX:6915639-6915859) is shown. Note that the entirety of chromosome 2L is shown on the left, while the X chromosome is shown on the right. **(d)** Putative Testis Enhancer 3 shows a proximal enrichment for H3K27ac signals in L3 larvae (modENCODE H3K27ac ChIP-Seq data, GSE49488). This region is located ~6.5kb upstream of Putative Testis Enhancer 3.

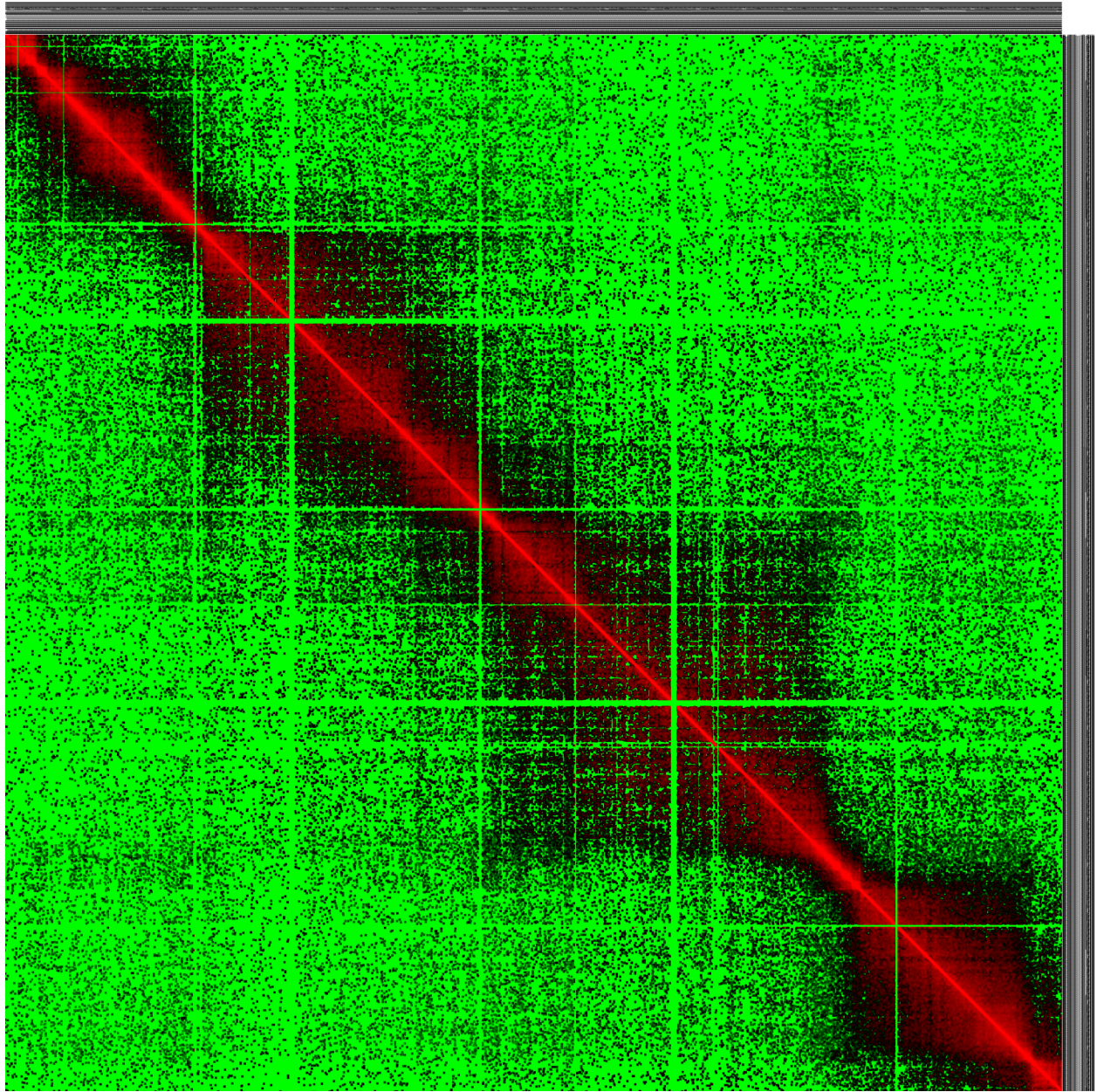


**Figure S4. FLEE1 is located within exonic sequence.** The FlyBase gene track and transcript view for FLEE1 shows that it is contained nearly entirely within the coding sequences of *MFS18* and *Elp3* on chromosome 2L.



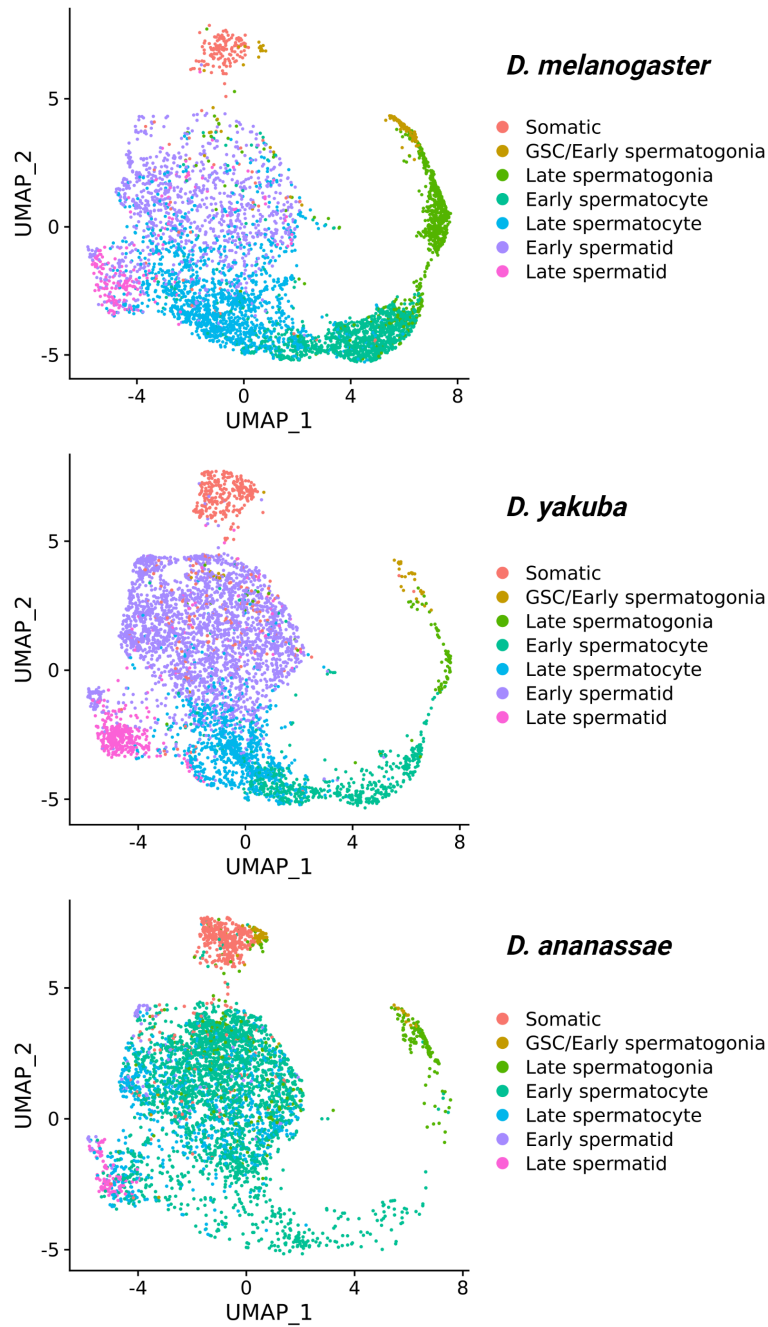


**Figure S5. Polymorphism and divergence in FLEE1/MFS18 locus.** Polymorphism and divergence calculations on chromosome 2L with a 2.5kb window, with the *MFS18* locus highlighted in green, show how the relative number of polymorphisms is low when compared to sequence divergence.



**Figure S6. Hi-C Structure for *D. pseudoobscura* larvae.** ~2Mb Hi-C structures of *D. pseudoobscura* chromosome 4 derived from L3 larvae and plotted at 1kb resolution. Window is identical to that in Figure 4e (green = low contact, black = medium contact, red = high contact).





**Figure S8. Unified UMAP projection across species.** Cell type assignments retrieved from (43) are displayed on a unified UMAP projection generated using Seurat, showing cell type assignments for *D. melanogaster*, *D. yakuba*, and *D. ananassae*.

## Supplementary Files

**Table S1. Co-expression and essentiality data for newly evolved distal duplicates.** Data for each new gene/parent gene pair with parental coexpression, maximal neighboring coexpression, and essentiality data reported.

**Table S2. Population genetic analysis of enhancer and HP6/Umbrea locus.** The test summaries of HKA, MK, and OBSM for positive selection of *MFS18* and *HP6/Umbrea*. “D, P, N, S, Obs, and Exp” indicate divergence, polymorphism, nonsynonymous sites, synonymous sites, observed numbers and proportions, and expected proportions, respectively.

**Movie S1. Live GFP expression.** Video of live larvae with FLEE1 under the control of an enhancer-reporter vector. Expression is seen restricted to the larval salivary glands. (MovieS1.mp4)

**Data S1. Sequence for FLEE1 enhancer.** Sequence for the cloned FLEE1 element in FASTA format.

>FLEE1

```
ACCAGGGCTTCGGTATGTTGCTGATGGAGGAGGCGGAGCGAATTGCTCGAGAGGAGCACGGCAGCAC
AAAACTGGCGGTCATATCGGGAGTGGGCACCAGAACTACTATCGCAAAATGGGATACCAACTTGAC
GGACCCTACATGTCAAAGAGCATAGAAGAAAATAACTAGGTATAGCGTTAAATGACTGTCTTGGTGGT
ATGTTGGAGGATTAATATTTGTATTTTATCACGCTAGGAGCAGTAAGATTTTCGCTACTTAAACTACT
CTCTTAAATATATACATTAATATATAGAATGAATCGATTTATTGGCTAAAACTCACAGGGTCCTTTAAA
GTATCAATGATACCACATATTTTTTGGCTTTAACATCTCACAAGAACAATAATGATCGTCAATCATA
AACGTGTATACTAAATAATATAAGCAGCATGAACATAAAATCGATCCACTCCAATATACCCACACATA
AATAAATAGGTTAGTTTTTCTGAGGAAAGTGTGCAAGAGAATGTTAAACGATGGCTTCCGCCGAACCA
AAGACTATAAATATGATCCAGCCAACCAAAATTGATGCCAGCAGCGGCGCTGAACACCATCGGCCAGC
TTTGTGTGAGCTCCAGAATGTGTCCGGCCAAGTATACTCCGAGAAAGCCAGGAATCGCGCCCACTGTG
TTCATCAGGCCAAAGACGCTGCCCCGAATGCAGAGGTGCCAGGTCTTGGGGATTCACTGTTACCGCGTT
GTTGTGGAAGCCCGTGCCGCCAATGATAATGGTCATGCAGATGAGCGCCGTATGGAAGTCCGAGGTG
CGGCTCATCACAAACAGGGCCAGATTCTGAGCGGCCAAAGCAGCAACTTTGGATGACCTTGCGCACCGT
CGTCGTGTGCCATTGCGGAGCGAGTAATCTGGTGGTCAAGTACTTGGCGAATAGCGTGCACGGTGGCA
GGGCAAGCCACGGGATCATGTTCACTACCCAACCCTTGGCGTGTGGAAAGCCGTGCGTGGAAAGTATGTA
GGCAGCCAGGAGAGTAGCACGAAGAAGCAGTTCATCTCGCAGGCGTGAGTCAGCACACAGGCCCAGA
AGGACAGCCTACGAAAGTATCGCAACCAAGGCACGGCTGACGTCTCTGCCGGACTCTTGTTTCGCGCAC
AGTCGGGATGGCGTGGCAATATTAATGATTCGGTTTTCGCTCGCCGGCCATTGCATAGTAGCGCAGCAC
CAGCGCCCATGCGATGCCCATCAGTCCTATCACCCGGAATACATACGACCAGCCGAAGTAGTCCAGCA
GAAAAGATCCCATAAATCCCAGTCAGAAGAGTACCTAGAGCCGATCCCGCTGTGAGCAGCCCAAAGAA
GCTGCTTCTCTCATTGGGGCACAAATTCTGCAAACGATTAAGTTATAGTTTATGTGTAAATTTATAAAA
TTAGCTAAGCACCTGACTGGTTAGACTAATCATGCTAGGAAAGTGCACGCCCTGAAGAGCGCCGTTCA
GGATTGCAATGGCAACAATGAAGGGAATAGCGTAACTCTTGATGGAGCCCGCCGTCCAGATGATAGT
GGGCATTAGGAATGTGATAAGCGACCAGCCGATTGCGGCAAAACAGAATGACTCGCTGGCCTCCAAAG
CGGTCGCTGAAGTAGCCGCCACAACTGCGTGAGTGTGTAGCCCCAGAAGAAGGAGCTGAGCACAG
TGCCCGAGTCGGTTTTGCTCCACTTTTGGGCGGATGCCACGGCCGGCACAGAAGTGGCATAGTGGTG
CGGGTGGAGTACAGCATACAGGTGCCCCGTAATAAGGGTGATGAACCAGACACGCTTCTCATGCCTGC
AATGATCCGACAAAGGAGTTGTAATTTGGGGAGGTTTAGTGAGCTGTATGCTGTGAGACCCACCTGGTC
CAAATGCTCTGCGTGTCCACCAGTTCCCCGCGCAGCAGAGAATATTTTAGCTTCTCGTCCATGGTCACA
AACTGGGTCCGGAATATTGCCTTTCTTCACGTACATATCAACTCCAATGCTTCGTTGCTTGCCGC
TGTGGCATATTTTACTGCCCTTTGTTTACTTTTACGTTGGCGACTAGACACGCCAAGTATTTGCG
CCTGTTAAAATTATGTTTTTACGTGGCCGTTTTTCCAACAGCCGCTGGACTAGAGCATAG
```