The Role of the 3-Dimensional Genome in New Gene Evolution

UnJin Lee[1,2*], **Deanna Arsala**[1], **Shengqian Xia**[1], **Mujahid Ali**[3], **Debora´ Sobreira**[4], **Ittai Eres**[4], **Qi Zhou**[3,5], and **Manyuan Long**[1*]

[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]MOE Laboratory of Biosystems Homeostasis & Protection, Life Sciences Institute, Zhejiang University, Hangzhou, Zhejiang, China
[*]correspondences to: ulee@mail.rockefeller.edu, mlong@uchicago.edu

ABSTRACT

In efforts to explain how duplicate gene copies may rise to fixation in a population, previous models of new gene origination have underappreciated the importance of the 3D genome in this process. We show that proximity-based regulatory recruitment in distally duplicated genes, i.e. enhancer capture, is an efficient mechanism for accommodation of new selective conditions. By performing a co-expression analysis on *D. melanogaster* tissue data and comparing essential to non-essential genes that have newly evolved, we show that enhancer capture is a significant driver of new gene evolution in distally duplicated genes. The new essential gene, HP6/Umbrea, is used as a model for understanding enhancer capture, as it evolved via a full duplication of the parental gene, its subsequent protein evolution is known, and it duplicated into a gene-poor region of the genome. HP6/Umbrea's expression pattern divergence from its parental gene, HP1b, as well as its high co-expression with neighboring genes suggest that it evolved via enhancer capture. ChIP-Seq data shows the presence of active enhancer marks appearing near HP6/Umbrea coinciding with onset of its expression which likely regulates HP6/Umbrea, its neighboring gene, as well as a distally located 6-gene cluster also found co-express with HP6/Umbrea. We find that these three loci, the putative enhancer, HP6/Umbrea, and the 6-gene cluster are in close physical proximity in the 3-D genome of *D. melanogaster*. Finally, we compare Hi-C data from two species with HP6/Umbrea, *D. melanogaster* and *D. yakuba*, to two species pre-dating HP6/Umbrea's insertion, *D. pseudoobscura* and *D. miranda*, showing that co-regulation of these same elements is the ancestral state and thus that HP6/Umbrea evolved via enhancer capture.

Keywords: new gene evolution, positional effect, enhancer

SIGNIFICANCE STATEMENT

Comprehensive analyses of new gene evolution across many clades have shown that the vast majority of new genes evolve via duplication-based methods, even in species with large population sizes. A few models have offered explanations for this seemingly paradoxical behavior, with the most commonly accepted ones being the duplication-divergence-complementation (DDC), escape-from-adaptive-conflict (EAC), and innovation-amplification-divergence (IAD) models. In this manuscript, we propose the enhancer-capture-divergence model of new duplicate gene evolution, where the rapid recombination of pre-existing protein-coding and regulatory elements offers the most efficient and evolvable path for modulating the protein production of an older gene. Subsequent to the fixation of this new variant, selection pressures are relaxed, e.g. through an environmental shift or the appearance of compensatory mutations elsewhere in the genome, allowing the new gene copy to begin to diverge in protein function. We provide genome-wide evidence for the enhancer-capture-divergence model using knock-down and expression data in *D. melanogaster*, while identifying the new essential gene HP6/Umbrea, a paralog of HP1b, as a model gene candidate for enhancer-capture-divergence.

1 INTRODUCTION

Genes arising from the class of duplication-based mechanisms are commonly inferred using synteny- and homology-based searches (Figure 1). While new genes are systemically understudied in comparison to their older counterparts, the most well-studied class of new genes are those originating from duplication-based mechanisms. However, in studying the evolutionary dynamics of duplication-based origination, a paradox arises: how do functionally redundant copies of the same gene rise to fixation?

The first models describing new gene evolution proposed that all new genes likely evolve via duplication-based mechanisms Ohno (1970). Under such a model, a duplicate copy of a gene is shielded from selective pressures, acquiring new mutations until a neo-functionalized copy of the gene provides sufficient selective force to carry this new gene to fixation. However, until advantageous function is acquired, the new gene copy is subject to genetic drift and is thus unlikely to rise to sufficient prevalence in a population to allow for rare, neo-functionalizing mutations to occur. This problem has been referred to as "Ohno's dilemma." Bergthorsson et al. (2007).

Various models have been proposed to resolve this problem - the duplication, divergence, complementation (DDC)/sub-functionalization model Force et al. (1999), the escape from adaptive conflict (EAC) model Hittinger and Carroll (2007), and the innovation, amplification, and divergence (IAD) model Bergthorsson et al. (2007); Nasvall et al. (2012) as well as its functional equivalent, the Adaptive Radiation model (AR) Francino (2005). However, these models fail to appreciate how the 3-dimensional genome and its corresponding regulatory landscape can drive neo-functionalization of a new gene from the moment of duplication (Figure 2).

To address how a duplicate, redundant gene copy may rise to fixation, these models all assume multiple functions for any studied gene. For pleiotropic genes, the DDC/sub-functionalization model allows for complementary non-functionalization of multiple functions that are originally shared between the duplicated copies (Figure 2a, b). Given a loss of function in one gene copy, the ability of the duplicate copy to compensate for this original loss of function confers a selective advantage to the duplicate copy. Eventually, under the DDC model, increasing divergence allows for the partitioning of multiple sub-functions between gene copies. Alternatively, while the DDC model allows each duplicate copy to possess only a subset of the original functions of the parental gene, the EAC model allows for increased optimization of multiple functions within the ancestral gene as each function partitions to each paralogous copy. Under this model, it may not be possible for a parental gene to simultaneously optimize each of its multiple functions. As such, duplication can allow for the relaxation of constraint on the evolution of the ancestral gene, thus resolving conflict and allowing for 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 fail to provide a mechanism for true neo-functionalization.

One common thread amongst the DDC and EAC models is their conformance to one of Kimura and Ohno's five governing principles of molecular evolution: "Gene duplication must always precede the emergence of a gene having a new function" Kimura and Ohta (1974). While complementary/optimizing mutations may stabilize the appearance of a duplicate gene copy, these mutations may only occur after the duplication of a new gene. The IAD model provides an alternative to this process by allowing for duplication itself to provide neo-functionalization via increased dosage for an auxiliary function of the original gene (Figure 2c). Here, the IAD model begins with an ecological shift that favors an auxiliary function of a gene, thus providing a selective advantage for high copy number. Importantly, as events of unequal crossing over are more common than point mutations, gene duplication occurs more frequently than substitutions and can thus fix in a population before regulatory changes evolve. Following this amplification, subsequent changes are accumulated on the various copies, allowing for divergence Bergthorsson et al. (2007).

While the IAD model provides a reasonable explanation for gene family expansions, particularly in the case of tandem duplications, some serious problems remain with the model, particularly when applied to multi-cellular organisms. While it is assumed that an ecological shift selects for higher copy number, it is not only the auxiliary function that is thus highly expressed, but the original function as well. The selective advantage conferred by increased dosage not only needs to be sufficiently greater than the metabolic costs of excess protein production, but it also needs to exceed potential deleterious effects caused by amplification of the original function. Depending on the spatio-temporal expression of the original gene, duplicate copies of the original gene will likely need to occur in a tissue-specific manner so as not to disrupt processes downstream of the original gene. Such precisely controlled expression is generally not of concern in single-celled organisms, where gene family expansions occur quite frequently. However indiscriminate expression of, for example, transcription factors within multi-cellular organisms will present a large selective barrier that copy number expansion must overcome, particularly if aberrations occur within key developmental processes.

One key factor missing in these models is the effect of chromosomal context on a new gene's regulatory function. A common thread amongst these various models is a separation of the initial establishment of a duplication followed by subsequent changes accumulated by various duplicate copies. Additionally, these models require that genes possess multiple functions. As an alternative to these models, we demonstrate that regulatory innovation via enhancer capture can also be a source of evolutionary novelty, allowing for rapid rewiring of gene regulatory networks in a single neo-functionalization step. During enhancer capture, neo-functionalization arises from the act of duplication itself by recombining pre-existing protein sequences with regulatory sequences, highlighting the importance of the three-dimensional eukaryotic genome in new gene evolution (Figure 2d).

2 RESULTS

2.1 Analysis of Tissue Co-Expression Shows 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 Bergthorsson et al. (2007); Nasvall et al. (2012). As previously described, one issue with this model is that there is an implicit assumption that during the environmental shift, the increase in fitness gained by over-activity of the auxiliary function must be greater than the decrease in fitness imparted by over-activity of the gene's original function(s). In the case of single-celled organisms where environments are encountered sequentially, it is reasonable to assume that selection might tolerate over-activity 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 multi-cellular organisms than in single-celled organisms, where a multi-cellular organism's overall phenotype is the cumulative (development) and simultaneous (organ systems) product of many different gene functions.

In the case of multi-cellular organisms, selection may increase for the expression of a gene within a single tissue type (Figure 2d). Under the IAD model, a full duplication will drive duplicate gene copies to fixation as it provides the most evolvable solution to new conditions. In contrast, under the enhancer capture-divergence model, a copy of the original gene duplicates into a region of the genome containing an active enhancer that increases expression in a tissue-specific manner. Alternatively, the new gene may migrate into a region of the genome containing unbound transcription factor binding sites, thus activating a pre-enhancer region into a new enhancer. Since the total output of the enhancer-capture-divergence model does not produce over-expression in other tissues like in the case of the IAD model, given sufficiently high population size, enhancer capture will be the more dominant mechanism for gene duplication, particularly with regards to distal/non-tandem duplications. This increase in fitness caused by the combined output of the new and parental genes thus drives both copies to fixation, providing an alternate resolution to Ohno's Dilemma. While enhancer capture remains the most rapid path to increasing fitness, compensatory mutations in the regulation of the parental gene may also provide a tissue-specific solution to increased selection. Once a compensatory mutation occurs, or even more simply, once the tissue-specific selection is relaxed, the new gene may then begin to diverge, accumulating substitutions.

Each model of gene duplication produces unique relationships between the expression patterns of a new gene vs its parent gene and/or a new gene vs its neighboring genes. As such, we may test whether enhancer capture drives the evolution of new genes evolving via distal/non-tandem duplication by utilizing tissue co-expression data. Specifically, we may predict to what degree a new gene will show tissue co-expression with its parental gene as well as with its neighboring gene depending on if the mechanism driving its evolution falls under the DDC, EAC, or enhancer-capture-divergence models.

Under the DDC or EAC models, the tissue expression patterns of parental and new genes are complimentary, resulting in low co-expression between parental and new gene copies ("parental co-expression"), while the tissue expression patterns of the new gene and its neighboring genes should have no relationship, resulting in random co-expression between the new gene and its neighboring gene co-expression ("neighboring co-expression"). Under enhancer capture, a broadly expressed parental gene acquires increased expression in select tissues by duplicating into a distant region of the genome under the control of an enhancer. 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 co-expression. On the other hand, since the new gene becomes regulated by the captured enhancer that is already influencing other genes, neighboring co-expression is high.

A tissue expression data set was obtained from FlyBase Larkin et al. (2021); Brown et al. (2014) (c.f. Methods and Materials) and co-expression between new/parental and new/neighboring gene pairs was calculated (Spearman correlation coefficient) for a set of new genes (N=87) which underwent a distal/non-tandem duplication of > 500kb whose essentiality has been validated experimentally Xia et al. (2021). This data contained tissue types extracted from both L3 larvae, pre-pupae, 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 (c.f. Methods and Materials). For tissues that were represented with multiple experimental runs, data from those tissue types were averaged prior to further analyses to avoid representation bias.

The resulting parent/neighbor co-expression plots ("PNC plot") for new essential genes (Figure 3a), new non-essential genes (Figure 3b), and both essential and non-essential genes (Figure 3c) can be used to test whether a significant number of distal/non-tandem duplications evolve via enhancer capture. We may define "low" and "high" co-expression as being below or above the median co-expression value across all distally duplicated new genes respectively. Genes that have evolved via enhancer capture should appear in the lower right quadrant in the PNC plots, as the expression patterns of the new gene diverges from the parental gene while the new gene and neighboring gene share the same expression pattern. Similarly, genes with that have evolved via the DDC or EAC models should appear in the bottom half of the PNC plots, with low parental co-expression resulting from divergent and complimentary expression patterns, and random neighboring co-expression as there is no expected relationship with the new gene and its neighboring genes.

Whiles genes in the lower right quadrant of the PNC plot may have evolved via the DDC/EAC models or enhancer capture, one key distinguishing feature of both models is how essential function is expected to partition between new gene and parental gene. Under the DDC/EAC models, all segregable functions of the original gene are expected to partition randomly between both parent and duplicate gene copies. As such, these models predict that essential gene function should also equally partition between both parent and new genes. The DDC/EAC models thus predict that the ratio of essential:non-essential genes in the entire lower half of the PNC plot, including the lower right quadrant, should match the overall ratio of essential:non-essential genes.

Alternatively, the enhancer-capture-divergence 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 (Supp. Figure S1). Specifically, selection for increased expression in a single tissue will result in elevated tissue-specific expression via the new gene copy, while all other function is retained in the parental copy, including its essential function; the new gene evolving via enhancer capture is expected to be non-essential while the parental gene is expected to be essential. As such, the enhancer-capture-divergence model predicts that the ratio of new essential:new non-essential genes in the lower right quadrant of the PNC plot should be significantly lower than the overall ratio of new essential:new non-essential genes (Table 1). Using the parent/neighbor co-expression plots, the ratio of new essential:new non-essential genes in the lower right quadrant (6:16) was found to be significantly lower than the overall ratio of new essential:new non-essential genes (35:52) using Fisher's Exact test (p=0.0256), suggesting that distally duplicated genes in *Drosophila melanogaster* primarily evolved via enhancer capture (Figure 3).

2.2 HP6/Umbrea as a Model for Enhancer-Capture-Divergence

While new genes categorically remain understudied, the evolution of HP6/Umbrea is a well-suited model system for understanding the enhancer-capture-divergence model as it is one of the few new genes whose protein evolution has been previously described in the literature (Figure 3, denoted as (*)) Ross et al. (2013). HP1b, a gene located on the X chromosome, duplicated approximately 12-15 million years ago (mya) into a gene-poor intronic region of dumpy, located on chromosome 2L (Figure 4). The new gene, HP6/Umbrea, was the result of a full duplication which included HP6/Umbrea's promoter region as well as its three known domains: the chromo domain, the chromo-shadow domain, and the hinge domain connecting the two.

Though HP6/Umbrea was lost ancestrally to multiple speciation events Ross et al. (2013), suggesting that the gene was not originally essential, HP6/Umbrea continued to evolve in a step-wise manner, diverging from its parental gene, HP1b. HP6/Umbrea subsequently lost its chromo domain approximately 10-12 mya; this was followed by an accumulation of key substitutions 0-7 mya, resulting in HP6/Umbrea's known essential protein function in *D. melanogaster* Greil et al. (2007); Chen et al. (2010); Xia et al. (2021). Using these results, protein neo-functionalization may be eliminated as the driving force behind the fixation of HP6/Umbrea given its step-wise protein evolution. Sub-functionalization and/or subsequent optimization of protein function may also be eliminated for similar reasons.

A simple comparison of HP6/Umbrea's expression pattern to the parental gene HP1b's very broad expression pattern suggests that HP1b is likely under the control of a simple constitutive-on promoter (Figure 4). Alternatively, while HP1b is found in all tissues, HP6/Umbrea is found only in a subset of tissues in which HP1b is found, suggesting that the duplication of HP1b's constitutive-on promoter into a region under control of an enhancer resulted in HP6/Umbrea's tissue expression pattern. This expression pattern is similar not to its neighboring gene, dumpy, but its second neighboring gene, CR44609, expressing primarily in the imaginal discs and male reproductive organs, demonstrating that these genes are likely co-regulated. Given that the tissue expression patterns of HP1b and HP6/Umbrea are not complimentary, sub-functionalization and/or subsequent optimization of regulatory function may also be eliminated as the driving force behind HP6/Umbrea's fixation.

In addition to results excluding other models, publicly available modENCODE ChIP-Seq/ChIPChip data Celniker et al. (2009) provides positive evidence that enhancer capture likely drove the early evolution of HP6/Umbrea. Using the embryonic S2 cell line as a negative control where there is little/no HP6/Umbrea expression, poised (H3K4me1) and primed (H3K27ac) enhancer marks in whole L3 larvae show strong enhancer activity in an intronic, gene-poor region of dumpy, coinciding with the onset of HP6/Umbrea transcription (Figure 4). Given the absence of other genes in the region (Figure 5a), HP6/Umbrea remains the likeliest target of the enhancer based on proximity and expression.

Given that it appears that HP6/Umbrea duplicated into a region that appears to be under the control of a pre-existing enhancer, we tested for further co-regulation in the region by using tissue expression data (c.f. Analysis of Tissue Co-Expression Shows New Genes Evolve by Enhancer Capture). We then applied a correlational analysis on this tissue expression data set to determine whether HP6/Umbrea is co-regulated with other neighboring genes. We took a 500kb region of the genome centered on the insertion site of HP6/Umbrea and calculated the tissue co-expression of each gene within this region. As enhancers function in a proximity-based manner, we would expect a distance-dependent effect on the co-expression of neighboring genes across the genome. To generate a baseline estimate of this distance dependent co-expression distribution, we sampled 1000 random genic loci within the *D. melanogaster genome*, calculating the degree of co-regulation expected on proximity alone. Notably, we find that using this distribution, the region of influence of any given regulatory region of the genome appears to be on the order of 25kb, suggesting that this is a characteristic distance for enhancer interaction in *D. melanogaster*. Outside of this region of influence, the likelihood of co-expression relaxes to the genomic average. Therefore, genes found within this region of influence with high tissue co-expression with neighboring genes are likely the result of co-regulation with the focal gene.

By comparing co-expression against this baseline distribution, we may find genes that share the same tissue-specific expression patterns as HP6/Umbrea and are thus likely co-regulated. As expected, we find that the neighboring gene, CR44609, possess the same expression pattern as HP6/Umbrea. Similarly, we find that a locus of 6 neighboring genes (CG11929, Elba3, CG3251, Taf12L, CG15631, CG42523) located approximately 100kb away from HP6/Umbrea also expresses in the same tissues as HP6/Umbrea, expressing primarily in the larval imaginal discs and male reproductive organs (Figure 5a).

While the co-expression of HP6/Umbrea's neighboring gene may be explained simply due to its proximity to HP6/Umbrea, the co-expression of the 6-gene cluster is not immediately evident as being a result of co-regulation. However, while this gene cluster is distally located along the chromosome beyond HP6/Umbrea's 25kb region of influence, due to the 3-dimensional nature of the eukaryotic genome, these genes may, in fact, be proximally located near HP6/Umbrea in 3D space and thus be co-regulated. Similarly, while active enhancer marks correlating to the onset of expression appears ∼50-100kb away from HP6/Umbrea, it is not immediately clear that these active enhancers are driving HP6/Umbrea expression, as its distance to HP6/Umbrea exceeds the 25kb region of influence. As the 3-dimensional conformations of the genome may still allow these distal genic elements to interact, we tested whether the putative larval enhancer, HP6/Umbrea, its neighboring gene, and the 6-gene cluster are co-regulated by examining high-resolution Hi-C data for *D. melanogaster* Wang et al. (2018) (Figure 5e, f). This data was aligned to the *D. melanogaster* genome dm6, and genome-to-genome contact frequencies were estimated using 5kb non-overlapping windows (c.f. Methods and Material).

Like co-expression, the frequency at which two genic elements make physical contact is expected to have a baseline, distance-dependent distribution. We may therefore test for co-regulation by predicting significant physical contact between HP6/Umbrea, its larval enhancer, and the cluster of co-expressed neighboring genes using Hi-C data in *D. melanogaster* (Supp. Figure S2). Such an interaction could be detected if contact between these two loci (i.e. HP6/Umbrea with enhancer and HP6/Umbrea with co-expressing genes) exceeds the baseline distance-dependent distribution of contact frequency. We generated an estimate of this baseline contact frequency distribution using 1000 independent loci that were sampled randomly from the genome, where contact data for the flanking regions were used to generate the baseline distance-dependent contact frequency distribution. We then extracted the contact frequency data for the HP6/Umbrea locus alone and compared this to the baseline genome-wide contact frequency distribution (Figure 5e, f).

We first note that after self-interactions are removed, we find that physical interactions in the genome generally remain highly localized, with most interactions lying near the focal locus as expected. Despite this, we find that HP6/Umbrea's complex contact distribution shows significant contact both with the putative larval enhancer as well as the neighboring 6-gene co-expression cluster (Figure 5e). Additionally, when this analysis is repeated for the 6-gene co-expression cluster, we find that this contact is reciprocated, as the 6-gene cluster shows significant contact across the cluster as well as with HP6/Umbrea (Figure 5g). Finally, HP6/Umbrea has enriched contact with the enhancer region that differentially activates at the onset of HP6/Umbrea expression. Combined with the tissue co-expression analysis, these results demonstrate that HP6/Umbrea and these 6 genes are likely co-regulated.

2.3 3D Genome Organization Pre-dates HP6/Umbrea Insertion

While we find evidence that HP6/Umbrea, the larval enhancer, and the 6-gene co-expression cluster are co-regulated, it is possible that these interactions evolved subsequent to HP6/ Umbrea's insertion. To determine whether these interactions pre-date HP6/Umbrea's insertion, we examined Hi-C data using a second in-group species, *D. yakuba* (shared by P. Reilly and P. Andolfatto, Supp Fig. S3), as well as newly generated data sets from two out-group species, *D. pseudoobscura* and *D. miranda* (Supp Fig. S4, S5) (Figure 5). While HP6/Umbrea inserted 12-15mya, the divergence between *D. melanogaster* and both outgroup species is 25mya Russo et al. (1995). Within these clades, *D. melanogaster* and *D. yakuba* diverged 6mya, while *D. pseudoobscura* and *D. miranda* diverged 4mya. While *D. melanogaster* Hi-C data was aligned to the standard reference genome (dm6), *D. yakuba*, *D. pseudoobscura* and *D.*

miranda were aligned to newer, high-quality reference genomes (*D. yakuba* shared by P. Reilly and P. Andolfatto, *D. miranda* from Mahajan et al. (2018), and newly generated *D. pseudoobscura*). In comparing the Hi-C contact patterns for both HP6/Umbrea and its neighboring co-expression cluster, we find that key features of the local chromosomal conformation are conserved: contact with the larval enhancer, reciprocal contact between HP6/Umbrea and its co-expression cluster and contact across the entire co-expression cluster (Figure 5d-e). The conservation of this chromosomal structure, even despite the subsequent evolution of protein function of HP6/Umbrea, suggests that the neo-functionalization event driving the fixation of the original duplication was likely driven by enhancer capture. Specifically, the 3D structure driving enhancer contacts existed prior to HP6/Umbrea's origination, and by duplicating into this region, HP6/Umbrea immediately captured this regulatory interaction.

3 DISCUSSION

3.1 Enhancer Capture Divergence Model

While various evolutionary mechanisms for the origination of new genes have been proposed, these models do not incorporate the 3-dimensional organization of the genome. In the DDC and EAC models, functions are sub-partitioned amongst paralogous copies, resulting in a neutral or adaptive process leading to the fixation of duplicate gene copies. However, in these models, subsequent substitutions in either gene copy are required to explain new gene origination, separating duplication from neofunctionalization. Alternatively, in the IAD model, duplication itself provides neo-functionalization by increasing dosage for an auxiliary function. In contrast to these models, we demonstrate how duplication itself may provide neo-functionalization in a tissue-specific manner, a result not predicted by these models. Such neo-functionalization provides a selective advantage in a direct, single-step mechanism without requiring subsequent substitutions as in the case of the DDC, EAC, and IAD models. In addition to producing gene fusions Wang et al. (2000) as well as favorable frame-shifts Wang et al. (2005), our model highlights the under-appreciated evolutionary value of both the act of duplication itself, and perhaps 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 Bridges (1936); Muller (1936), the advent of new chromosomal conformation capture technologies allows us to directly connect the conservation of chromosomal domains Harmston et al. (2017); Krefting et al. (2018) and the origination of new genes under a strong conceptual framework.

Under the enhancer-capture-divergence model, a gene copy duplicates into a pre-existing regulatory context (Figure 6a), gaining a new regulatory interaction. Alternatively, the duplication may occur in a region of the genome possessing transcription factor binding sites (pre-enhancer) but isn't yet acting as an active enhancer due to a paucity of nearby genes to regulate. Regardless of exact mechanism, due to the 3-dimensional looping nature of the eukaryotic genome, duplication recombines genes and enhancers into new combinations, thus resulting in regulatory novelty (Figure 6b, c). As such, this model provides an explanation and mechanism for the well-described but poorly-understood phenomenon where new genes often possess highly tissue-specific expression patterns Zhang et al. (2019); Dai et al. (2006); Vibranovski et al. (2012); Long et al. (2013a) (Supp. Figure S1). Here, selection for increased expression in a single tissue is most rapidly achieved by acquiring a new tissue-specific expression pattern via distal duplication. Subsequently, the appearance of compensatory mutations near the parental gene or the relaxation of selective pressures allows for this newly fixed duplicate gene copy to begin to diverge, resulting in either inactivation or the gain of new protein function.

The enhancer-capture-divergence model also provides a mechanistic explanation by which gene interaction networks may rapidly evolve Zhang et al. (2015). Under this model, we have two separate gene interaction sub-networks for both parental and neighboring genes (Figure 6d). 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 downstream function of the original parental gene's protein function (Figure 6e) while simultaneously preserving the pre-existing interactions from both parental and neighboring genes' sub-networks. As the act of duplication is more likely to occur than a point mutation Bergthorsson et al. (2007); Nasvall et al. (2012), enhancer capture will therefore be a faster route to generating increased tissue-specific expression of a parental gene (Figure 1) than any set of mutations in the parental gene's regulatory sequence. As a consequence, these new genes can fix, allowing for the subsequent accumulation of substitutions. While duplications occur more frequently than substitutions, point mutations altering the regulation of the parental gene will continue to occur. If eventually a compensatory mutation in the parent gene allows for increased tissue-specific expression, this will then allow the new gene to be free from the pressures of natural selection and thus evolve further, resulting either in pseudogenization, e.g. as in the case of HP6/Umbrea's loss in *D. eugracilis*, or the acquisition of further function, e.g. as in the case of HP6/Umbrea's gain of essential function in *D. melanogaster* Ross et al. (2013).

One key aspect of the enhancer-capture-divergence model is the selective advantage imparted by increased tissue-specific expression. While the EAC model describes a very narrow enhancer-based explanation for gene duplication and fixation Hittinger and Carroll (2007), the resolution of evolutionary conflict, such as sexual antagonism, is a well-known driver of the evolution new genes Kursel et al. (2021); VanKuren and Long (2018). While most new genes have highly tissue-specific expression patterns, these often favor either the female or male reproductive organs/germlines in *D. melanogaster* Long et al. (2013a). A close examination of the expression pattern of HP6/Umbrea demonstrates the same – HP6/Umbrea is expressed primarily in the imaginal discs and the male reproductive organs. Similarly, the parental gene HP1b appears to have expression highly skewed towards the female reproductive organs. As such, 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 wide-spread mechanism for the resolution of such sexual antagonism, providing a rapid, one-step mechanism for acquiring differential expression between sexes.

Central to both the enhancer-capture-divergence and IAD models is the rapidity at which novelty is produced. Such rapid evolvability arguments may provide an explanation for the origination of the eukaryotic genome, organized into multiple chromosomal domains that result in a segregation of regulatory enhancer sequences and protein-coding genic sequences. While our model is illustrated with different tissue types, we may easily substitute various environmental conditions for tissue type. Under the context of sequential environmental conditions, the amplification of auxiliary function during transient environments is sensible as described by IAD model Bergthorsson et al. (2007), as precise spatio-temporal regulation of the original gene is no longer needed, assuming that environmental conditions return back to "normal." Crucially, paralogs become fixed in the IAD model as duplication is the most evolvable solution to altered selective demands. As plasticity arises when permanent genic solutions are not easily evolvable Lee et al. (2022) or when future environmental conditions are completely unpredictable Skanata and Kussel (2016), duplications into genomic regions where enhancers already exist may produce precise epigenetic control of a given protein much more rapidly than divergence via accumulated substitutions. Furthermore, while epigenetic mechanisms exist in prokaryotic genomes, these remain simple binary switches as in the case of the lac operon Jacob and Monod (1961). As the number of environmental conditions increase, the requisite gene-network complexity for such regulation becomes a large barrier for further evolution. Co-regulation of multiple genic units is already an efficient and useful method for dealing with multiple environmental conditions as demonstrated by the lac operon. By developing enhancers that operate in a proximity-based manner, eukaryotic genomes thus provide for the expansion of co-regulation into modular structures Wagner (1996) capable of handling greater than two distinct conditions without the need for developing three-way (or larger) switches. Given that enhancer capture can accelerate evolution both through faster-than-substitution alterations as well as modularity, the eukaryotic genome's inherently higher evolvability may suggest that enhancer capture may be one clue in understanding the evolutionary origins of the nucleus.

3.2 Revisiting an Old Theory of New Genes

Current models of eukaryotic gene regulation roughly defines two broad classes of genomic sequences: protein-coding sequences and regulatory sequences Ong and Corces (2011). Under these models, the precise spatio-temporal control of a protein-coding sequence is provided by genomic enhancer elements where the concerted binding of transcription factors acts to either increase or decrease the activation energy of transcription. Importantly, such control occurs in a three-dimensional, distance-dependent manner enhancer elements may only control genomic elements that are physically close to these enhancers within the eukaryotic nucleus Ong and Corces (2011). Due to this proximity-based effect, the exact conformation of the genome is significantly more important in understanding gene regulation than simple gene order, particularly in gene-dense genomes. Using this proximity-based effect, we show that the chromosomal context into which a gene duplicates, particularly non-tandem/distal duplications, may generate novel enhancer-gene interactions that immediately neo-functionalizes duplicate gene copies.

These positional effects have been well-described since the origins of the field of genetics. The first known positional effect was described in the study of the *bar* gene in *Drosophila melanogaster* in 1925 by Alfred Sturtevant a mere 12 years after he developed the first genetic map Sturtevant (1925). In his original allelomorphic series, Sturtevant surmised that a duplication must have occurred with the *bar* gene, where two copies of the gene were inherited along a single chromosome. Crucially, in comparing the homozygous *B/B* phenotype to the *BB/B*$^-$ phenotype, Sturtevant found that the *double-bar* or *ultrabar* allelomorph produced a more extreme phenotype than expected by dosage alone. This *double-bar* or *ultrabar* allelomorph of the classic *bar* gene was found to be the result of a gene duplication event through the examination of polytene chromosomes by Calvin Bridges in 1936 Bridges (1936). Dobzhansky recognized this as what he called a positional effect and that it was a result of some kind of chromosomal interaction with neighboring genes Bridges (1936). Soon afterwards, Hermann Muller recognized the importance of this observation for the origination of new genes:

> *“We consider the point of chief interest in the Bar case to be its illustration of the manner of origination of extra genes in evolution. Bar had for a long time offered the best case yet known for the idea that genes could arise de novo*. Its interpretation as some sort of duplication met with difficulty, in our ignorance of the real existence of a ’position effect’...”*
>
> -Hermann Muller (Science, 1936)

*note ”*de novo*” is not used indicate a particular new gene origination mechanism as in Long et al. (2013b).

4 METHODS AND MATERIALS

4.1 Tissue expression data and analysis

Tissue expression data was retrieved from FlyBase. Pre-computed RPKM data files were downloaded, with RPKM values for each FlyBase transcript being reported for 29 tissues Brown et al. (2014). As many of the tissues types were repetitive, data from head, ovary, carcass, and digestive system were averaged to reduce over-representation bias in further correlational analyses. Gene map data was also obtained from FlyBase to properly identify neighboring genes Larkin et al. (2021). Parental/new gene pair information was retrieved from Chen et al. (2010). Spearman correlation coefficients were calculated using the tissue expression data between parental and new gene pairs. Due to 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. Additionally, correlation coefficients for all genes within 500kb of HP6/Umbrea were reported. To generate a baseline distancedependent genomic estimate of co-expression, 1000 random genic loci were chosen and co-expression values (Spearman) between the randomly selected gene and all neighbors within a 500kb range were calculated. This 500kb region was then divided into 100 non-overlapping windows where mean and variance in correlation coefficients was calculated across all randomly selected loci.

4.2 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 Celniker et al. (2009). H3K4me1 ChIP-Chip data for S2-DRSC cells was obtained using data ID 304 and 3760. H3K27ac ChIP-Chip data for S2-DRSC cells was obtained using data ID 296 and 3757. H3K4me1 ChIP-Seq data for whole Oregon-R L3 larvae was obtained using data ID 4986. H3K27ac ChIP-Seq data for whole Oregon-R L3 larvae was obtained using data ID 5084. For all data sets, data was obtained in .gff3 format and visualized using the UCSC Genome Browser.

4.3 Hi-C data

Publicly available Hi-C libraries were obtained from NCBI: *D. melanogaster*, PRJNA393992. *D. yakuba* Hi-C data was shared by Patrick Reilly and Peter Andolfatto, and *D. pseudoobscura* and *D. miranda* data was shared by Mujahid Ali and Qi Zhou. *D. melanogaster* source tissue was S2 cells, *D. yakuba* from adult females, and *D. pseudoobscura* and *D. miranda* were L3 larvae. Hi-C libraries were preprocessed, mapped, and filtered using HiCUP version 0.8.0 Wingett et al. (2015). Specifically, reads from fastq files were trimmed at ligation junctions, and subsequently each mate of paired-end sequences were independently mapped to the respective genomes using bowtie2 version 2.2.9 Langmead and Salzberg (2012). Reads were mapped to genomes consisting of canonical chromosomes only (i.e. excluding scaffolds and other unplaced sequences). *D. melanogaster* reference genome was dm6 and obtained from FlyBase Larkin et al. (2021). The *D. yakuba* reference genome was shared by Patrick Reilley and can be obtained from NCBI (PRJNA310215). The *D. pseudoobscura* reference genome was obtained directly from Ryan Bracewell (https://www.ryanbracewell.com/data.html) Bracewell et al. (2019) and the *D. miranda* reference genome was obtained from NCBI (PRJNA474939), Mahajan et al. (2018). HiCUP was used further to remove experimental artifacts based on an *in silico* genome digest as previously described Wingett et al. (2015). HiCUP mapped and filtered .sam files were then converted to formats compatible with HOMER version 4.11 Heinz et al. (2018) and juicer tools version 1.22.01 Durand et al. (2016). To create matrices, HOMER was used to tile the genome into matrices of fixed-size bins, and assign reads to their correct intersecting bins. HOMER was also used to normalize contact counts in these matrices based on known Hi-C biases, as previously described Heinz et al. (2018). Juicer tools was used to produce .hic files at resolutions of 5kb for *D. melanogaster* and *D. yakuba* and 7.5kb for *D. pseudoobscura* and *D. miranda*, and to create normalized matrices.

Using Hi-C contact matrices, data rows for HP6/Umbrea and its neighboring cluster were pulled for a 400kb region centered on HP6/Umbrea and self-self interactions were removed. To generate a genome-wide distance-dependent distribution of contact, 1000 random loci were sampled. Contact data for each locus was then normalized with total contact (arb. units) being equal for all loci. The mean and variance for each non-overlapping window was calculated and reported and compared to HP6/Umbrea and the co-expression clusters' data. To generate genomic coordinates for HP6/Umbrea before duplication, *D. melanogaster* sequence flanking HP6/Umbrea's insertion site was aligned to the *D. yakuba*, *D. pseudoobscura* and *D. miranda* reference genomes using blast. Similarly, the promoter region of CG11929 was aligned to *D. yakuba*, *D. pseudoobscura* and *D. miranda* reference genomes to represent the co-expression cluster.

5 ACKNOWLEDGEMENTS

U.L. is funded in part by GM7197-42. M.L. was supported by the National Institutes of Health grant 1R01GM116113-01A1. D.A. is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number F32GM146423. This manuscript is dedicated in memory of Frances Lee.

REFERENCES

Bergthorsson, U., Andersson, D. I., and Roth, J. R. (2007). Ohno's dilemma: Evolution of new genes under continuous selection. *Proc. Natl. Acad. Sci. USA*, 104:17004–17009.

Bracewell, R., Chatla, K., Nalley, M. J., and Bachtrog, B. (2019). Dynamic turnover of centromeres drives karyotype evolution in drosophila. *eLife*, Sep 16;8:e49002.

Bridges, C. B. (1936). The bar ”gene” a duplication. *Science*, 83:210–211.

Brown, J., Boley, N., Eisman, R., May, G., Stoiber, M., Duff, M., Booth, B., Wen, J., Park, S., Suzuki, A., Wan, K., Yu, C., Zhang, D., Carlson, J., Cherbas, L., Eads, B., Miller, D., Mockaitis, K., Roberts, J., Davis, C., Frise, E., Hammonds, A., Olson, S., Shenker, S., Sturgill, D., Samsonova, A., Weiszmann, R., Robinson, G., Hernandez, J., Andrews, J., Bickel, P., Carninci, P., Cherbas, P., Gingeras, T., Hoskins, R., Kaufman, T., Lai, E., Oliver, B., Perrimon, N., Graveley, B., and Celniker, S. (2014). Diversity and dynamics of the drosophila transcriptome. *Nature*, 512:393–399.

Celniker, S. E., Dillon, L. A. L., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., Kellis, M., Lai, E. C., Lieb, J. D., MacAlpine, D. M., Micklem, G., Piano, F., Snyder, M., Stein, L., White, K. P., Waterston, R. H., and modENCODE Consortium (2009). Unlocking the secrets of the genome. *Nature*, 459:927–930.

Chen, S., Zhang, Y. E., and Long, M. (2010). New genes in drosophila quickly become essential. *Science*, 330:1682–1685.

Dai, H., Yoshimatsu, T. F., and Long, M. (2006). Retrogene movement within- and between-chromosomes in the evolution of drosophila genomes. *Gene*, 385:96–102.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., and Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell Systems*, 3.

Force, A., Lynch, M. F., Pickett, B., Amores, A., Yan, Y., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151:1531–1545.

Francino, M. P. (2005). An adaptive radiation model for the origin of new gene functions. *Nature Genetics*, 37:573–577.

Greil, F., de Wit, E., Bussemaker, H. J., and van Steensel, B. (2007). Hp1 controls genomic targeting of four novel heterochromatin proteins in drosophila. *EMBO Journal*, 26.

Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Markenschlager, M., and Lenhard, B. (2017). Topologically associating domains are ancient features that coincide with metazoan clusters of extreme noncoding conservation. *Nature Communications*, 8:441.

Heinz, S., Taxari, L., Hayes, M. G. B., Urbanowski, M., Chang, M. W., Givarkes, N., Rialdi, A., White, K. M., Albrecht, R. A., Pache, L., Marazzi, I., Garcia-Sastre, A., Shaw, M. L., and Benner, C. (2018). Transcription elongation can affect genome 3d structure. *Cell*, 174:1522–1536.

Hittinger, C. T. and Carroll, S. B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, 449:677–681.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3:318–356.

Kimura, M. and Ohta, T. (1974). On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. USA*, 71:2848–2852.

Krefting, J., Andrade-Navarro, M. A., and Ibn-Salen, J. (2018). Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biology*, page 87.

Kursel, L. E., McConnel, H., de la Cruz, A. F. A., and Malik, H. S. (2021). Gametic specialization of centromeric histone paralogs in drosophila virilis. *Life Science Alliance*, 4:e202000992.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9:357–359.

Larkin, A., Marygold, S., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P., Goodman, J., Gramates, L., Millburn, G., Strelets, V., Tabone, C., Thurmond, J., and Consortium, T. F. (2021). Flybase: updates to the drosophila melanogaster knowledge base. *Nucleic Acids Research*, 49:D899–D907.

Lee, U., Mortola, E., Kim, E., and Long, M. (2022). Evolution and maintenance of phenotypic plasticity. Biosystems 222 (2022): 104791.

Long, M., VanKuren, N. W., Chen, S., and Vibranovski, M. D. (2013a). New gene evolution: little did we know. *Annual Review of Genetics*, 47:307–333.

Long, M., VanKuren, N. W., Chen, S., and Vibranovski, M. D. (2013b). New gene evolution: little did we know. 47:307–333.

Mahajan, S., Wie, K. H.-C., Nalley, M. J., Gibilisco, L., and Bachtrog, D. (2018). De novo assembly of a young drosophila y chromosome using single-molecule sequencing and chromatin conformation capture. *PLoS Biology*, 16:e2006348.

Muller, H. J. (1936). Bar duplication. *Science*, 83:528–530.

Nasvall, J., Sun, L., Roth, J. R., and Andersson, D. I. (2012). Real-time evolution of new genes by innovation, amplification, and divergence. *Science*, 338:384–387.

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer, New York.

Ong, C.-T. and Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 12:283–293.

Ross, B. D., Rosin, L., Thomae, A. W., Hiatt, M. A., Vermaak, D., de la Cruz, A. F. A., Imhof, A., Mellone, B. G., and Malik, H. S. (2013). Stepwise evolution of essential centromere function in a drosophila neogene. *Science*, 240:1211–1214.

Russo, C. A., Takezaki, N., and Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. *Molecular Biology and Evolution*, 12:391–404.

Skanata, A. and Kussel, E. (2016). Evolutionary phase transitions in random environments. *Physical Review Letters*, 117:038104.

Sturtevant, A. H. (1925). The effects of unequal crossing over at the bar locus in drosophila. *Genetics*, 10:117–147.

VanKuren, N. W. and Long, M. (2018). Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nature Ecology and Evolution*, 2:705–712.

Vibranovski, M. D., Zhang, Y. E., Kemkemer, C., Lopes, H. F., Karr, T. L., and Long, M. (2012). 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 Biology*, 10:49.

Wagner, A. (1996). Does evolutionary plasticity evolve? *Evolution*, 50:1008–1023.

Wang, Q., Sun, Q., Czajkowsky, D. M., and Shao, Z. (2018). Sub-kb hi-c in d. melanogaster reveals conserved characteristics of tads between insect and mammalian cells. *Nature Communications*, 9:188.

Wang, W., Zhang, J., Alvarez, C., Llopart, A., and Long, M. (2000). The origin of the jingwei gene and the complex modular structure of its parental gene, yellow emperor, in drosophila melanogaster. *Molecular Biology and Evolution*, 17:1294–1301.

Wang, W., Zheng, H., Yang, S., Haijing, Y., Li, J., Jiang, H., Su, J., Yang, L., Zhang, J., McDermott, J., Samudrala, R., Wang, J., Yang, H., Yun, J., Kristiansen, K., Wong, G. K. S., and Wang, J. (2005). Origin and evolution of new exons in rodents. *Genome Research*, 15:1258–1264.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., and Andrews, S. (2015). Hicup: pipeline for mapping and processing hi-c data. *F1000Res*, 4:1310.

Xia, S., VanKuren, N. W., Chen, C., Zhang, L., Kemkemer, C., Shao, Y., Jia, H., Lee, U., Advani, A. S., Gschwent, A., Vibranovski, M. D., Chen, S., Zhang, Y. E., and Long, M. (2021). Genomic analyses of new genes and their phenotypic effects reveal rapid evolution of essential functions in drosophila development. *PLoS Genetics*, 17:e1009654.

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwent, A. R., Yu, Y., Hou, G., Zi, J., Zhou, R., Wen, B., Zhang, J., Chougule, K., Wang, M., Copetti, D., Peng, Z., Zhang, C., Zhang, Y., Ouyang, Y., Wing, R. A., Liu, S., and Long, M. (2019). Rapid evolution of protein diversity by de novo origination in oryza. *Nature Ecology and Evolution*, 3:679–690.

Zhang, W., Landback, P., Gschwend, A. R., Shen, B., and Long, M. (2015). New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biology*, 16:202.

Model	Parent/New Gene Co-expression	New/Neighboring Gene Co-expression	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

Table 1. Model Summary. Different models of new-gene evolution are compared in the context of distal/ectopically duplicated genes, the duplication-divergence-complimentation model (DDC), the escape-from-adaptive-conflict model (EAC), the innovation-amplification-divergence model (IAD), and the enhancer-capture-divergence model. Each model predicts different relationships between pairs of parental and new genes, as well between both new gene and its neighboring gene. Genes likely driven by each model can be found in their respective locations in the parent/neighbor-gene co-expression (PNC) plots. Additionally, the segregation of essential function in these genes is assessed.

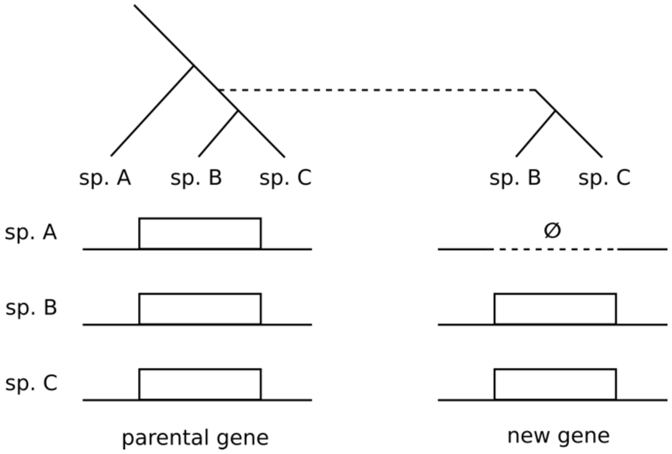

Figure 1. Identification of new genes. The insertion of a new gene may be inferred by using syntenic alignments of closely related species. Gaps within these alignments may be used to determine the location of a new gene insertion, while reciprocal-best searches may determine whether a gene arose via duplication as well as the identity of the parental gene.

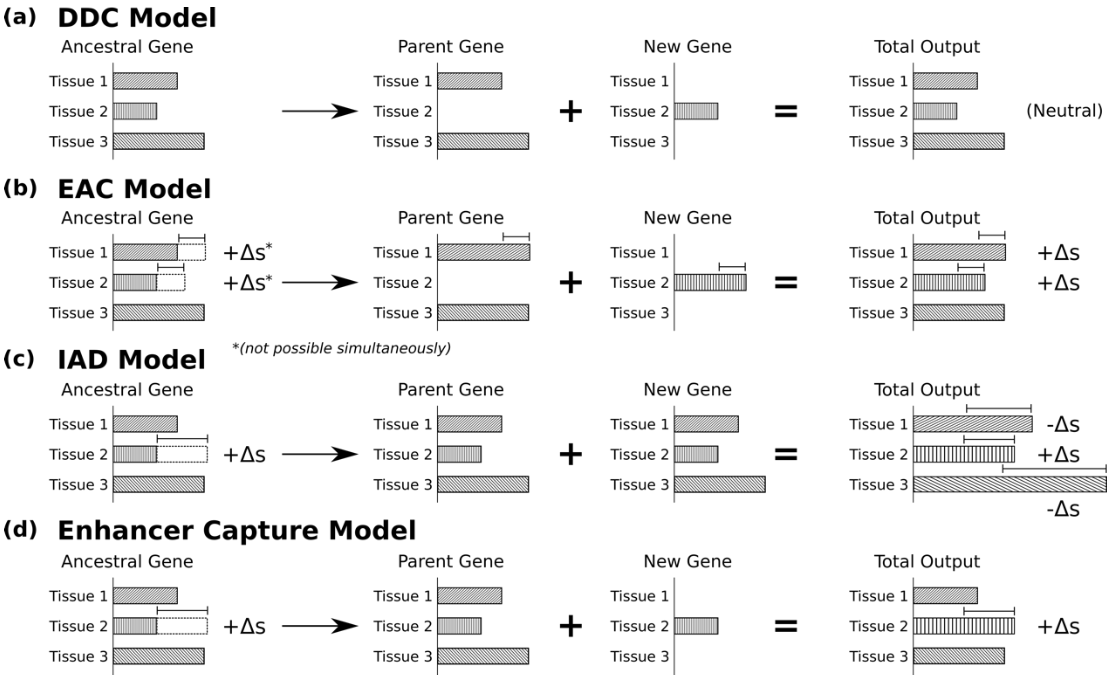

Figure 2. Comparison of extant models. Various evolutionary models have been proposed to explain how redundant gene copies become fixed in populations ("Ohno's Dilemma"). Presented are illustrations for the (a) Duplication-Divergence-Complementation (DDC), (b) Escape-from-Adaptive-Conflict (EAC), (c) Innovation-Amplification-Divergence (IAD), and (d) enhancer-capture-divergence models, where the gene regulation of three tissue types are considered and optimal conditions are shown in dotted boxes. Under the DDC model (a), redundancy allows for compensation of any single loss-of-function event, eventually causing the expression pattern of the original gene to be segregated between both parent and new genes. Given that the original protein is produced by both the parent and new genes, the total output is identical to the original gene, and is thus a neutrally evolving process. Under the EAC model (b), two functions cannot be optimized within a single gene copy, and this conflict is resolved via the act of duplication, allowing for simultaneous optimization of both copies. The total output of these two gene copies now has higher fitness than the output of the original gene, rising to fixation. Under the IAD model (c), an environmental shift causes increased selection for an auxiliary function of the original gene. As duplication events (unequal crossing-over) occur more frequently than point mutations, duplication of the original gene provides a more rapid accommodation of the new environmental conditions than regulatory mutations by increasing dosage. However, while this model allows for increased fitness due to increased auxiliary function, one issue in this model is that this increase in fitness must also overcome the penalty imposed by over-activity of all other functions. Over-activity is generally not an issue when environments change sequentially, as is the case of single-celled organisms, but incorrect regulation can be a significant barrier in multi-cellular organisms, e.g. in the case of key transcription factors. Under the enhancer-capture-divergence model (d), increased expression of a single function provides a selective advantage. A region of the genome contains an enhancer/pre-enhancer that increases fitness once a gene copy duplicates into a region under its control (thus activating it in the case of a pre-enhancer). As the original protein is produced by both parent and new gene, the total output of both parent and new gene increases overall fitness, thus driving both copies to fixation.

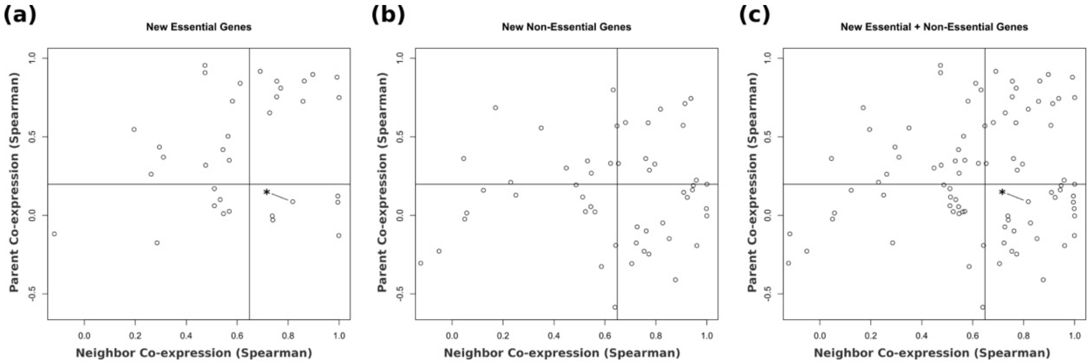

Figure 3. New genes evolve via enhancer capture. Shown are parent/neighbor tissue co-expression patterns for new genes in *D. melanogaster* which have migrated either more than 500kb away or between chromosomes (new essential genes (a), new non-essential genes (b), and combined essential & non-essential genes (c). Tissue co-expression (Spearman correlation coefficient) between new gene/parental gene pairs is plotted on the vertical axis while maximal tissue co-expression between new gene/neighboring genes pairs is plotted on the horizontal axis. Note, the co-expression between the new gene and four of its neighbors was calculated, two on each side, and the maximal co-expression is reported here. Vertical and horizontal lines indicate median co-expression value of all distally duplicated new genes as in (c). Genes which evolved via enhancer capture are expected to have low parental co-expression and high neighboring co-expression and should thus be present in the lower right quadrant. Genes evolving under the DDC or EAC models should have low parental co-expression due to complimentary expression patterns and random neighboring co-expression. While a new gene's essential function is equally likely to be partitioned between either parent or new gene under the DDC or EAC models, new genes evolving via enhancer capture are unlikely to have essential function, as the expression of the new gene will only augment existing expression of the parental gene, leaving the original essential function intact. Comparing the overall ratio of new essential to new non-essential genes (35:52) to the ratio of new essential to new non-essential genes showing high neighboring/low parental co-expression (6:16) shows that new genes evolve via regulatory capture (Fisher's Exact, p=0.0256). (* denotes HP6/Umbrea.)

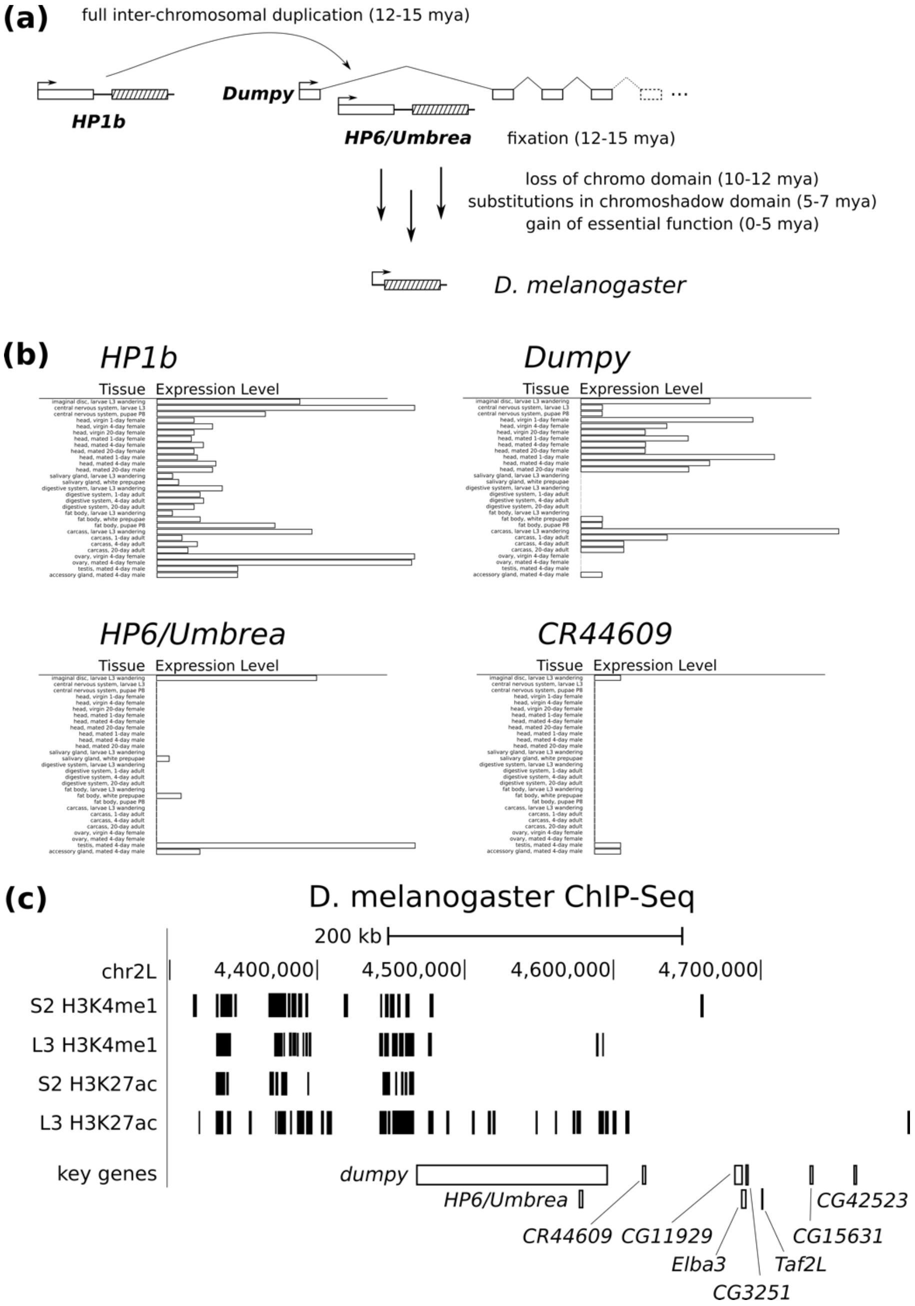

Figure 4. HP6/Umbrea evolved via enhancer capture. (a) HP6/Umbrea is a new essential gene in *D. melanogaster* which arose from a full duplication of HP1b into an intronic region of dumpy, migrating from chromosome X to 2L. HP6/Umbrea's well characterized, step-wise protein evolution suggests that amino-acid substitutions were unlikely to have driven the duplicate gene copy to fixation. (b) Unlike the broad expression pattern of 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 and male reproductive organs. This expression pattern is shared with HP6/Umbrea's neighboring gene CR44609. (c) A comparison of ChIP-Seq 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 100kb intronic region of dumpy that is, aside from HP6/Umbrea, devoid of protein coding genes.

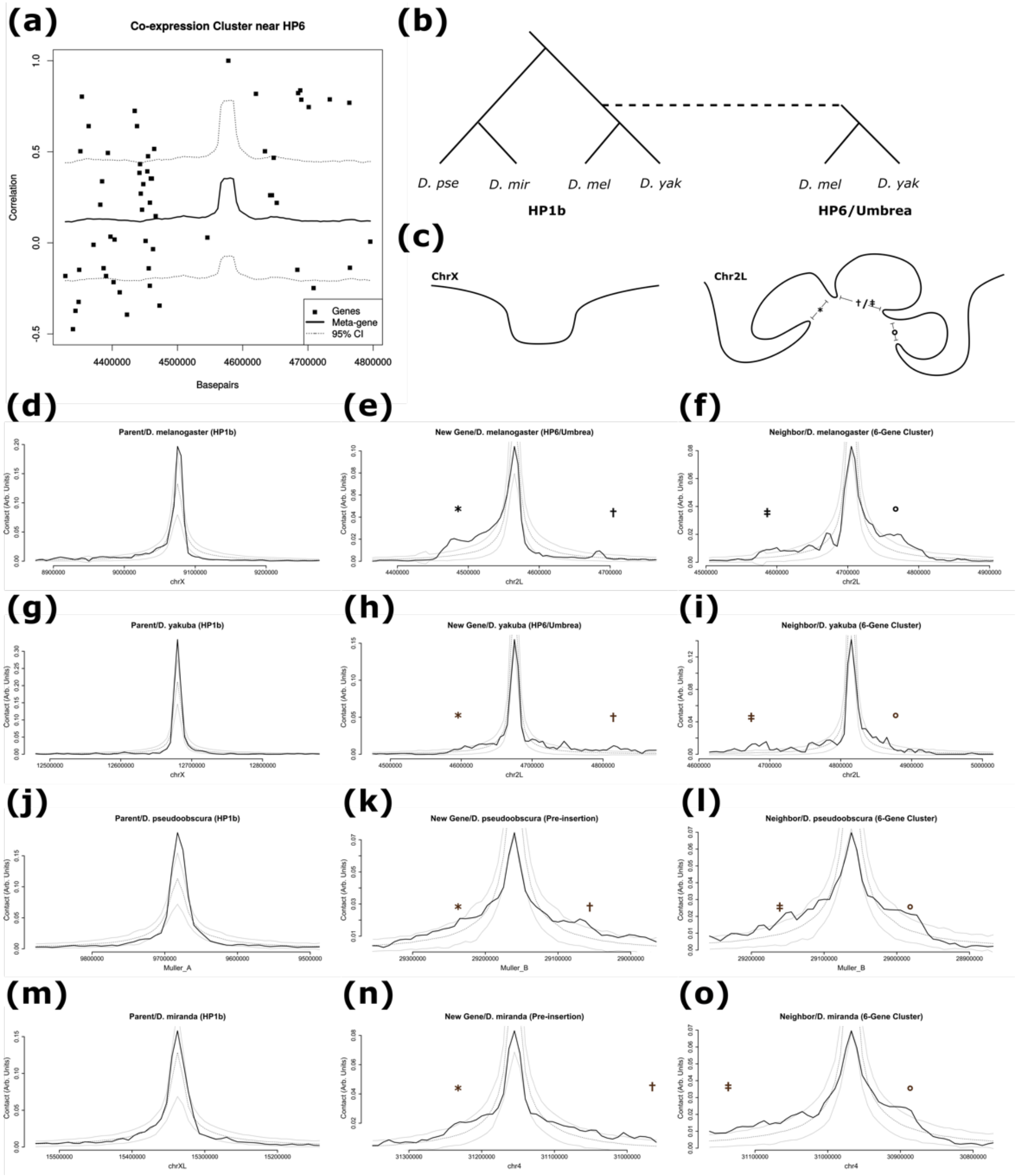

Figure 5. HP6/Umbrea co-expression is associated with conserved chromosomal looping that pre-dates its insertion. (b) Tissue co-expression analysis between HP6/Umbrea and neighboring genes reveals the presence of a co-regulated cluster of 6 neighboring genes. Note absence of other genes within dumpy's intronic regions. (b) Two in-group species, *D. melanogaster* and *D. yakuba* (div. ~ 6mya), contain HP6/Umbrea, while two out-group species, *D. pseudoobscura* and *D. miranda* (pse-mir div. ~ 4mya, pse-mel div. ~ 25mya), pre-date HP6/Umbrea's insertion (~ 12-15mya).(c) Cartoon legend illustrating features in (d)-(o). Not drawn to scale. (d)-(o) Hi-C data tracks for in-group (*D. mel* (d)-(f), *D. yak* (g)-(i)) and out-group (*D. pse* (j)-(l), *D. mir* (m)-(o)) species are shown for the parental gene HP1b (left column) HP6/Umbrea's insertion site (middle column) and the co-regulated 6-gene cluster (right column), with a 95% confidence interval generated from genomic sampling plotted in dotted lines. On the vertical axis is contact in arbitrary units, and on the horizontal axis is genomic coordinates centered on the viewpoint location. Conserved feature (*) shows that HP6/Umbrea's insertion site loops with the active larval enhancers contained in dumpy's intronic gene-desert. Conserved features (†) & (‡) show that HP6/Umbrea's insertion site reciprocally loops with the co-regulated 6-gene cluster. Conserved feature (◦) shows that the co-regulated gene cluster loops across the entire 6-gene cluster.

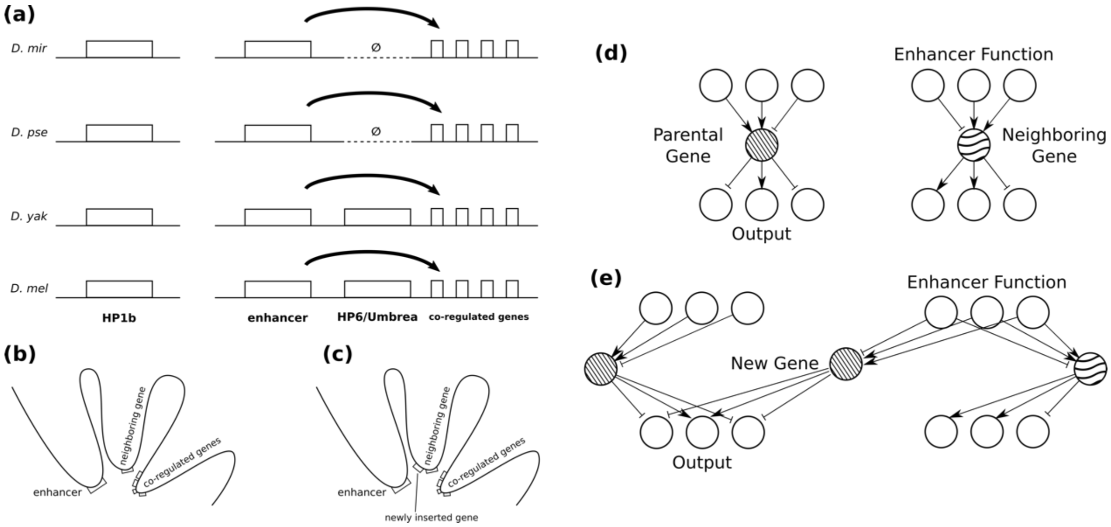

Figure 6. The 3D organization of the genome allows for rapid rearrangement of genetic networks. Panel (a) depicts a cartoon illustration of the action of the larval enhancer on the neighboring cluster of co-regulated genes as well as the future insertion site of HP6/Umbrea. (b) Preceding insertion of HP6/Umbrea, the larval enhancer was in contact with both HP6/Umbrea's neighboring gene as well as with the co-regulated 6-gene cluster. (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 (d) pre- & (e) post-duplication, is depicted above. Note that parental gene and neighboring gene's original interactions remain intact, preserving previous function.

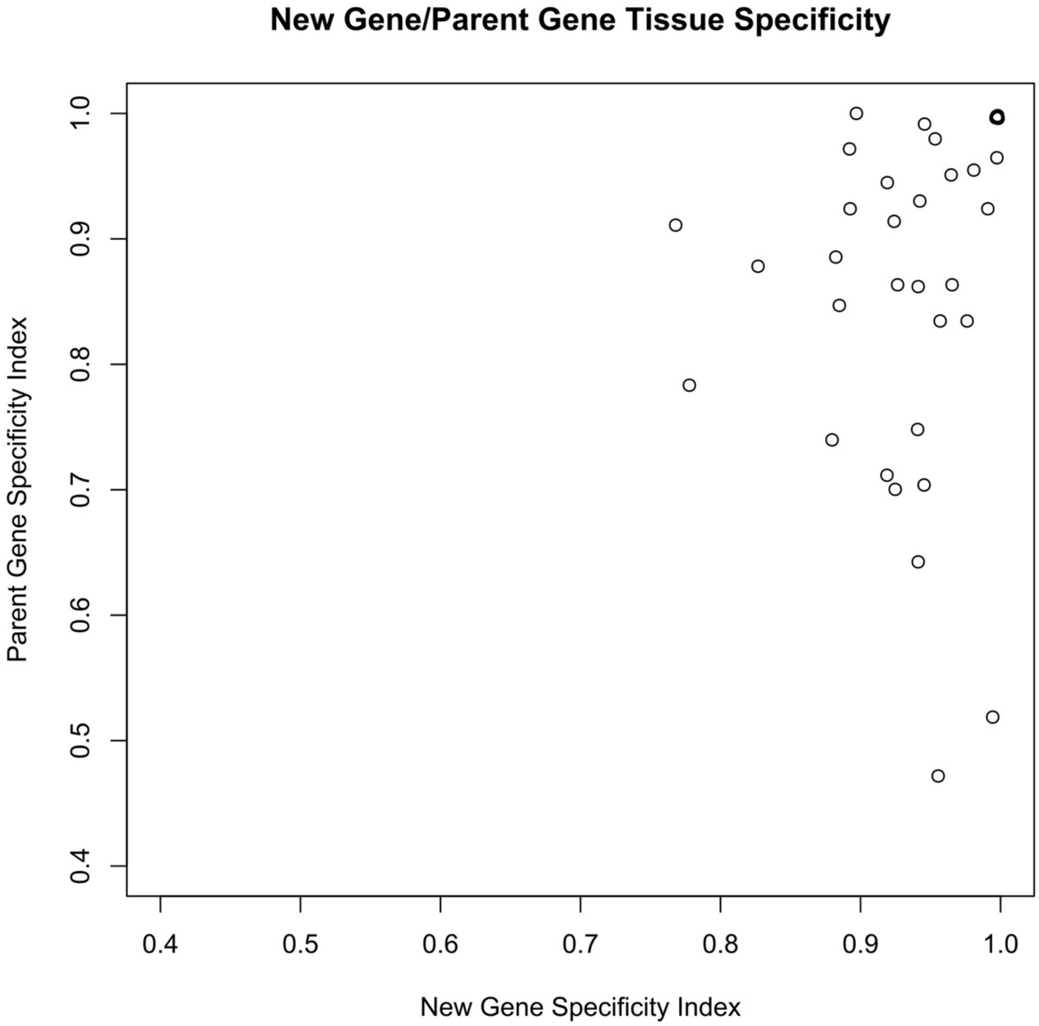

Figure S1. New genes evolving via distal/ectopic duplication *D. melanogaster* demonstrate higher tissue specificity than parental genes. Using new-gene/parent-gene pairs for genes evolving via distal/ectopic duplication in *D. melanogaster*, the tissue specificity index *tau* is calculated and plotted above, demonstrating that new genes evolving via ectopic/distal duplication have higher tissue specificity than parental genes.

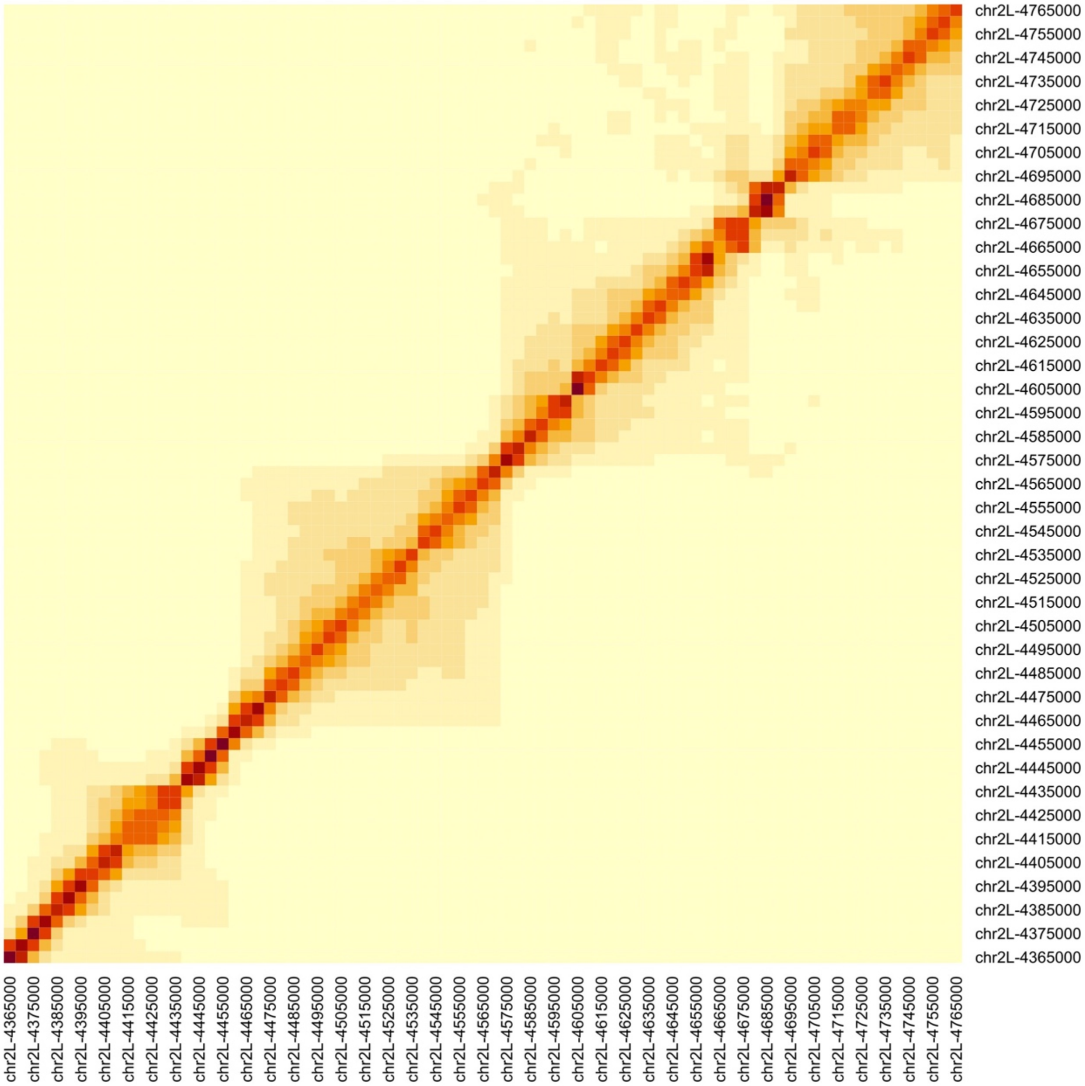

Figure S2. Local Hi-C heatmap for *D. melanogaster.* Shown above is the local chromosomal configuration of chromosome 2L in the vicinity of HP6/Umbrea (chr2L:4570000) and the neighboring co-expression cluster (chr2L:4710000).

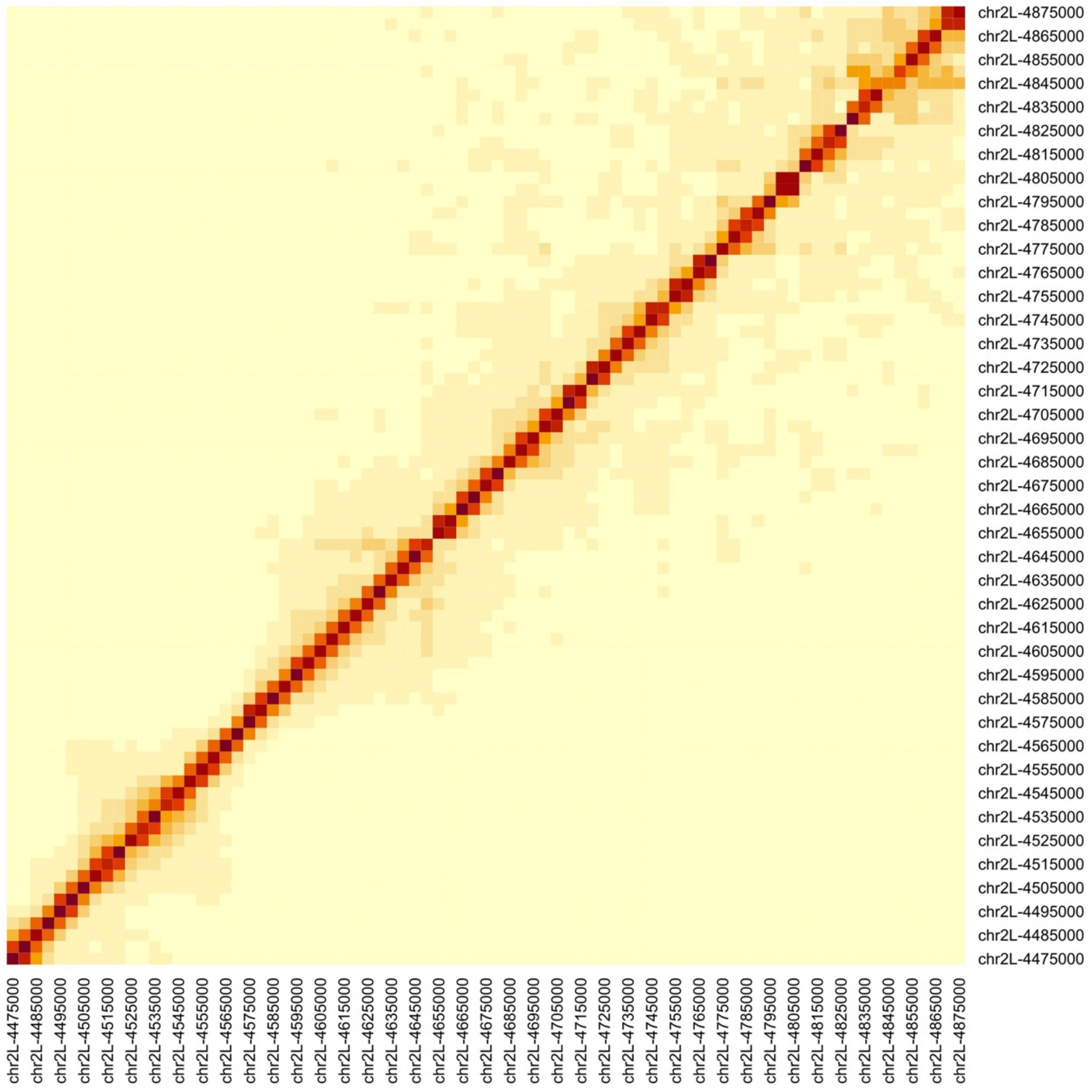

Figure S3. Local Hi-C heatmap for *D. yakuba*. Shown above is the local chromosomal configuration of chromosome 2L in the vicinity of HP6/Umbrea (chr2L:4680000) and the neighboring co-expression cluster (chr2L:4820000).

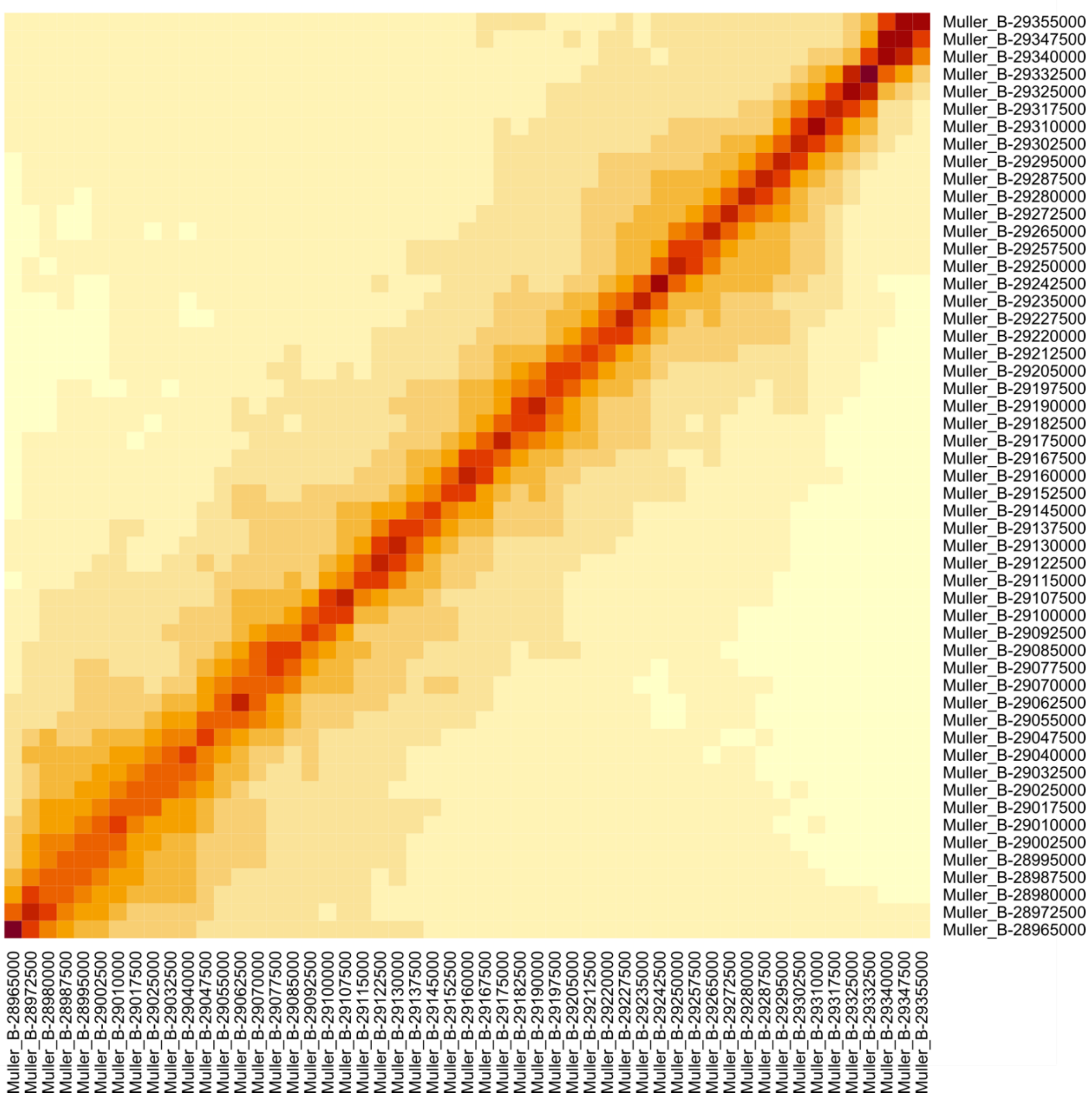

Figure S4. Local Hi-C heatmap for *D. pseudoobscura*. Shown above is the local chromosomal configuration of Muller Element B in the vicinity of HP6/Umbrea's future insertion site (Muller B:29165000) and the neighboring co-expression cluster (Muller B:29070000).

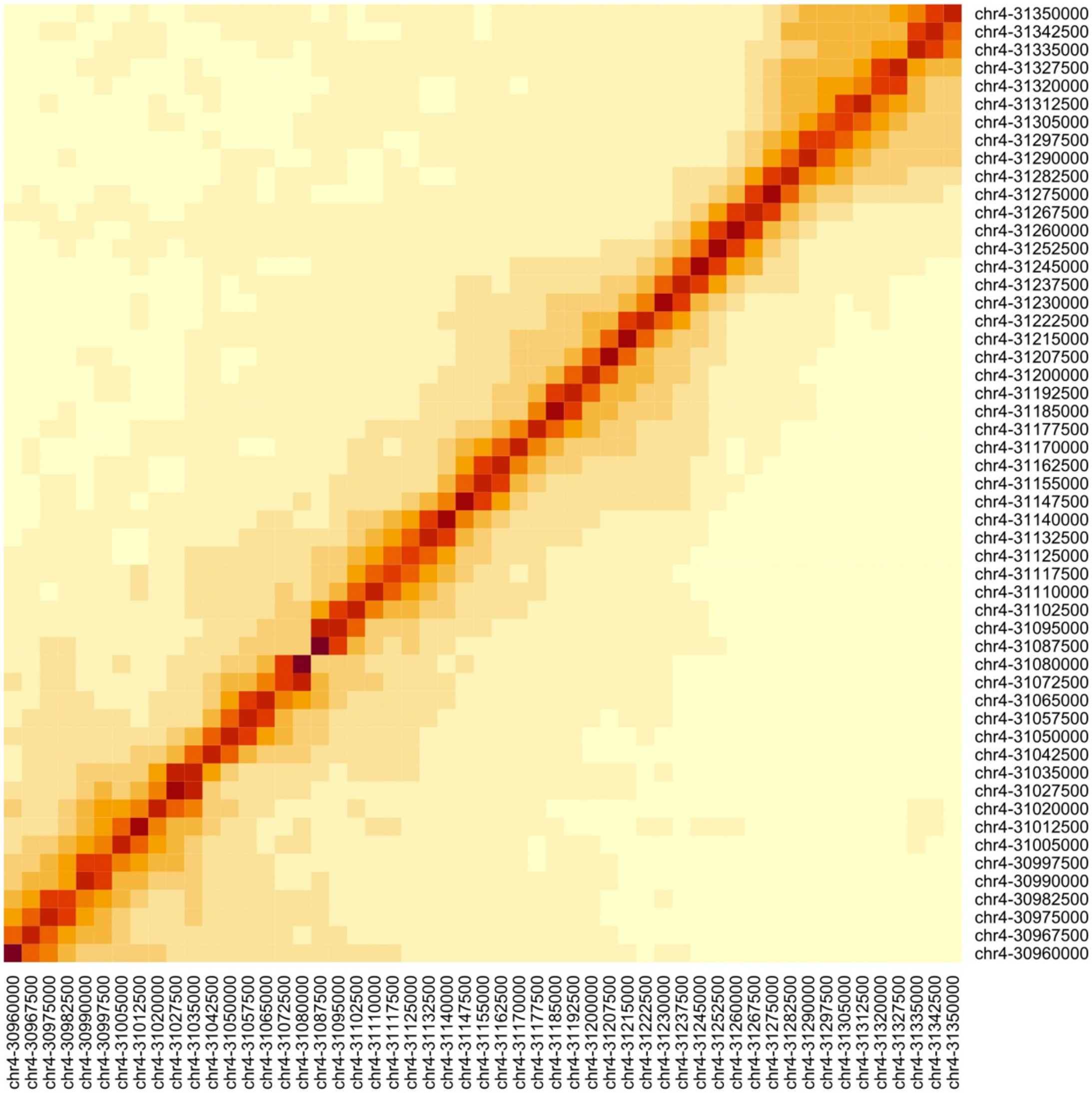

Figure S5. Local Hi-C heatmap for *D. miranda*. Shown above is the local chromosomal configuration of chromosome 4 in the vicinity of HP6/Umbrea's future insertion site (chr4:31160000) and the neighboring co-expression cluster (chr4:30970000).