The Pennsylvania State University

The J.Jeffrey and Ann Marie Fox Graduate School

EVOLUTION AND EXPRESSION OF OVERLAPPING GENES IN DROSOPHILA

A Thesis in

Molecular Cellular and Integrative Biosciences

by

Luyi Wo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2024

The thesis of Luyi Wo was reviewed and approved by the following:

Stephen W. Schaeffer
Emeritus Professor of Biology
Thesis Advisor

Kateryna Makova
Professor of Biology
Verne M. Willaman Chair of Life Sciences

Melissa Rolls
Paul Berg Professor of Biochemistry
Head of the Molecular Cellular and Integrative Biosciences Graduate Program

ABSTRACT

In *D. melanogaster*, over 30% of the protein coding genes overlap and yet we are not clear about the evolution and regulation of overlapping genes. Using the 12 *Drosophila* (fruit fly) genome sequences and strand specific sequencing data in *D. melanogaster*, we found they are generally more conserved than non-overlapping genes and the conservation degree correlates with overlapping configuration, namely, relative orientation and overlapping proportion. This conservation might relate to their transcriptional regulation during development and are not due to functional relationship between the partners. Different overlapping configurations also show different evolutionary patterns with gene gains and losses being the dominant mechanisms of change rather than rearrangements. Therefore, we showed that overlapping genes have distinctive expression and evolutionary patterns compared to physically close gene clusters and they are not a neutral phenomenon, which supports the non-random gene distribution hypothesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my extreme gratefulness to my thesis advisor Dr. Stephen Schaeffer, who helped me build the solid foundation for genomic and evolutionary research. He introduced me to fundamental skills such as Perl programming and spreadsheet techniques, familiarized me with database and professional tools, suggested courses to take and books and literature to read, supported me going to professional conferences and workshops and fixed my English speaking and writing. He himself is an exemplary biologist full of love and passion for his research, which enlightens me through daily communication and guidance. He is extremely patient and always encourages me and gives lots of critical comments during this project. After years of halting of this project, he is still willing to offer his advisership on this thesis even after his retirement. Without his kindness and professional suggestions, this thesis wouldn't have been possible.

I would like to thank my committee member and program chair Dr. Melissa Rolls, who helped me transfer to the master program in MCIBS in 2015. When I reached out to her for a possible resume of study, she kindly offered me an extension to finish my thesis. Without her administrative coordination, this thesis wouldn't have been possible. I would like to thank my committee member Dr. Kateryna Makova, who used to be chair of my doctoral committee. I learned basic concepts of molecular evolution from her which laid the theoretical foundation of this work. And she gave lots of insightful feedback on my research during my initial study. It is a great honor for me to have her back on my thesis committee.

In addition, I would like to thank Dr. Richard Ordway and Dr. Hong Ma for recruiting me to the graduate program in Genetics in 2009. I want to thank Dr. Kimberlyn Nelson who showed me the nitty-gritty of how to teach undergraduate labs, which also consolidated my molecular lab skills along the way. I want to thank my former committee members Dr. Anton Nekrutenko, Dr.

Claude dePamphilis, Dr. Yu Zhang who gave me helpful feedback on my research. I would like to thank my friends from the statistical department who have already graduated from Penn State by now. Dr. Lejia Lou, Dr. Wanghuan Chu helped me learn statistical theories and Dr. Yihan Li inspired me on the expression analysis of this study.

I deeply appreciate this precious opportunity to be able to present my work and hopefully add a little knowledge to the scientific community. I owe gratitude to everyone who has helped and supported me in various ways along this special journey. This project was partially funded by National Science Foundation Grant No. NSF OCI- 0904166 (S. W. Schaeffer). I also received financial support through teaching assistantships from the Biology Department and first year fellowship award from the graduate program in Genetics.

Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Chapter 1

Introduction

1.1 A brief review of protein coding gene structure and function of components in eukaryotes.

A typical eukaryotic protein coding gene contains exons and introns. The transcription start site and stop site mark the boundaries of a typical gene. The transcripts of a gene may then go through a series of modifications, such as intron splicing, 5' capping and polyA tail addition to produce mature mRNA for translation. The first exon contains a 5' UTR region, which is a segment between transcription start site and start codon and does not code for protein. The same is true for the last exon, which contains a segment between the stop codon and transcription stop site, called the 3' UTR. Therefore, a protein coding gene contains two kinds of sequences, the coding sequence, which is both transcribed and translated, and the sequence that is only transcribed but not translated, namely, introns and UTR regions. Although non-coding sequences such as intron and UTRs do not directly contribute to the coding information of the genome, they play important roles in the regulation of gene expression. There are cis-regulatory regions within the intergenic DNA sequence, which controls and regulates gene expression. There are promoter regions locating immediately upstream of the first exon and there are enhancers, silencers and insulators which are not necessary located near the gene but function through formation of DNA loops with the help of trans-regulatory proteins.

1.2 The chromatin architecture behind the distribution and coregulation of genes

Genes are not distributed randomly on the chromosome. Only 10% of fly genes are located relatively solitary in the genome whereas most genes form gene clusters (Li et al., 2010).

With the era of genomic sequencing in the early 2000s highly expressed house-keeping gene clusters were found in humans (Lercher et al., 2002), in flies (Spellman & Rubin, 2002; Thygesen & Zwinderman, 2005) and in zebra fish (Tsai et al., 2009). The mechanism behind this non-random distribution and co-regulation of gene clusters is not clear. Recent advances in chromatin architecture research have begun to shed some light on this question.

To pack large DNA molecules on the scale of meters into the small eukaryotic nucleus on the scale of micrometers requires delicate and well-regulated mechanisms (Fraser et al., 2015). Cremer & Cremer (2001) demonstrated that DNA molecule is nonrandomly organized in the nucleus forming chromosomal territories. Histones have long been known to help with folding of DNA strand into small nucleosomes and compressed to the scale of hundreds of base pairs. They form a spool-like structure that the DNA strand wraps around, and different modifications of histones correlates with the active or inactive activity states for the transcription of wrapped DNA strand (Kharchenko et al., 2011). The chromatin with small bundles of nucleosomes then goes through further folding and looping to form topologically associated domains (TAD). The median size of TADs in Drosophila was initially estimated to be 108kb (Hou et al., 2012; Sexton et al., 2012) but later estimated to be 26 kb (Eagen et al., 2017; Ramírez et al., 2018) as techniques such as Hi-C with larger sample sizes increased resolution over the years. TADs with active genes and inactive genes are then packed separately into two compartments: transcriptionally active open compartment A and closed inactive compartment B (Lieberman-Aiden et al., 2009). This 3D topology of chromatin can bring linearly distant DNA elements together, thus rendering long range regulation of gene expression such as the binding of enhancers with promoters possible. It has been reported active gene rich regions are within the range of so-called transcription factories where the transcription machinery such as transcription factors, RNA polymerase, enzymes are in high concentration and readily available to use, which would facilitate the high expression level of genes within (Fraser et al., 2015). Further, recent studies have shown a relative paucity of gene

rearrangement events within TADs versus their boundaries (Liao et al., 2021; Wright & Schaeffer, 2022). Moreover, TAD structure is conserved to a certain degree across various tissues (Schmitt et al., 2016) and across developmental stages (Hug et al., 2017) and corresponds well with the conserved orthologous gene syntenic blocks (Liao et al., 2021). Together, these findings suggest genes within the same TAD structure may be co-evolved and co-regulated.

1.3 Overlapping genes.

The discovery of overlapping genes was rather surprising (Barrell et al., 1976). If two genes reside at the same location on the DNA molecule, they would have complete identical or complementary DNA sequence to each other, which is unprecedented compared with typical non-overlapping genes. In this scenario, one mutation in the coding sequence overlap could result in sequence changes in two protein coding genes. Even a mutation in non-coding sequence overlaps might also affect the regulation of the two genes. Moreover, overlapping genes are the extreme case of spatial proximity, meaning any DNA level biochemical activity such as gene transcription and translation would affect both genes. However, overlapping genes have been found to be prevalent in viral, bacterial, mitochondrial and eukaryotic genomes (Behura & Severson, 2013; Chen & Stein, 2006; Chirico et al., 2010; Johnson & Chisholm, 2004; Makalowska et al., 2005; Pavesi, 2006; Pavesi et al., 2018; Rogozin et al., 2002; Sakharkar et al., 2005) Several studies based on eukaryotic genomes found that around 4-10% of genes overlap (Table **1-3-1**). In addition, overlapping genes were found to have quite complex configurations given the complexity of gene structure in eukaryotic genomes. A consensus of literature is to divide overlapping genes into subgroups based on their sequence context which could result in different biochemical and evolutionary consequences (Table **1-3-2**). Given the possible sequence constraints and prevalence of overlapping genes, it is thus natural to ask the following questions 1) how common are overlapping genes in each species and whether overlapping genes are under

selection due to sequence and physical constraints; 2) how do overlapping structures occur and how do they evolve; and 3) whether there is transcriptional correlation between overlapping genes pairs. The literature review in this section is centered around current findings regarding these three questions.

Table **1-3-1**: Overlapping genes in various eukaryotic genomes.

Species	**Genome size**	**Number of genes**	**Gene Density**	**Number of overlapping genes**	**Percentage**	**References**
Human	3.3 Gb	22291	1/148042bp	2978	13.36%	Makalowska et al., 2005
Chimpanzee	3.8 Gb	21506	1/176695bp	2219	10.32%	
Mouse	2.8 Gb	25383	1/110310bp	3456	13.62%	
Rat	2.9 Gb	22159	1/130872bp	1080	4.87%	
Chicken	1.2 Gb	17709	1/67762bp	1960	11.07%	
Fugu	400 Mb	20796	1/19234bp	993	4.77%	
Zebrafish	1.5 Gb	23524	1/63765bp	1625	6.91%	
A. aegypti	1376 Mb	14613	1/94163bp	1522	10.42%	Behura & Severson, 2013
A. gambiae	278 Mb	12519	1/22206bp	1130	9.03%	
C. elegans	100 Mb	19765	1/5059bp	2380	12.04%	Chen & Stein, 2006

Table **1-3-2**: Different overlapping types.

		Opposite Strand Overlap	Same Strand Overlap
Transcripts Overlap	**Partial**	**Convergent or Tail-to-tail** **Divergent or Head-to-head**	**Tandem or Parallel**
	Full	**Nested-host Antisense or Parent-embedded Antiparallel**	**Nested-host Sense or Parent-embedded Parallel**
CDS Overlap	**Phase 0**	1 2 3 **1 2 3** 3 2 **1 3 2** 1	**1 2 3** 1 2 3 1 2 3 **1 2 3**
	Phase 1	1 2 3 **1 2 3** **3 2 1** 3 2 1	**1 2 3** 1 2 3 1 2 **3 1 2** 3
	Phase 2	1 2 3 **1 2 3** 3 **2 1 3** 2 1	**1 2 3** 1 2 3 1 **2 3 1** 2 3

1) Overlapping genes in viruses and prokaryotes are thought to be selectively advantageous in organisms where small genome sizes are favored due to physical space constraints, cell cycle replication constraints, or regulation coordination. In viruses, the number of overlapping genes is negatively correlated with genome size (Belshaw et al., 2007), supporting a compact genome hypothesis (Sakharkar et al., 2005). But the genome size hypothesis cannot explain overlapping genes in eukaryotes. Eukaryotic genomes are comprised of highly repetitive, moderately repetitive, and unique sequence DNA. Among *Drosophila* species, the unique genic sequences within the euchromatin appear to be a uniform fraction of the genome, but these genomes have expanded their fractions of repetitive DNA especially in heterochromatic regions near centromeres (Schaeffer et al., 2008), which are shown mainly composed of retroelements

(Chang et al., 2019). These data suggest the sizes of eukaryotic genomes are not constrained. Therefore, alternative models are needed to study the overlapping genes in eukaryotic genomes.

Generally, two lines of evidence are adopted to infer whether overlapping genes in eukaryotic genomes are under selection. One way is to compare the percentage of ortholog retention in closely related species. If overlapping structures are often found to be retained across the studied species more than chance effect, one would conclude they are under purifying selection. The other way is to compare sequences between overlapping orthologs of closely related species and perform statistical tests on the Ka/Ks ratio. Tajima's *D* could be compared between genomic backgrounds to reveal whether overlapping orthologs are under selection. However, there is no consensus of conservation level of overlapping genes. Case studies of overlapping genes show that the level of conversation differs case by case (Bukhnikashvili, 2023; Hudson et al., 2007; Munroe et al., 2015). A study based on the *Gnas* and *Gnal* loci investigated a complex region with multiple overlapping genes and multiple configurations and found large differences in the evolutionary rate among overlapping loci, with some showing complete conservation while others showed little shared structure (Wadhawan et al., 2008). On the genomic level, studies treating overlapping genes as a whole group were found not to be conserved in human and mouse genomes (Veeramachaneni et al., 2004), among other vertebrate species (Makałowska et al., 2007), or across metazoan (Soldà et al., 2008). However, by further partitioning of overlapping genes, Chen & Stein (2006) concluded that overlapping genes are generally conserved in *C. elegans* and its closely related species. While nested overlapping genes were shown to be generally under weaker selection in humans (Yu et al., 2005) and in flies (Lee & Chang, 2013), Assis et al. (2008) instead pointed out nested genes are a neutral phenomenon where the non-coding intronic regions of existing genes can tolerate the gain of new genes without affecting parent genes in which they are embedded, thus leading to a more complex genome with even more nested structures over the long term. These discrepancies among

different species are hard to form a consensus due to vast differences between divergence times of different taxa used to infer conservation degree. For instance, comparative studies based on *C. elegans* and *C. briggsae* which diverged around 39 Myr ago found a higher amount of conserved orthologs than studies based on human and mice which diverged 87 years ago (Kumar et al., 2017). Also, the pooling of overlapping gene samples and mixing of different types might have averaged out the different selection signals from different configuration of overlaps. Therefore, a detailed comparative analysis discriminating overlapping types and including more species on various divergence times is needed for assessing the conservation degree and rate of evolution of overlapping genes.

2) Many mechanisms have been proposed for the origin of overlapping genes such as overprinting, gene duplication and rearrangements, and exon extension of existing genes (Calvete et al., 2012; Keese & Gibbs, 1992; Makalowska et al., 2005; Wright et al., 2022). A study based on 3'-UTR overlapping genes ACAT2 and TCP1 in human showed that they were brought together by one chromosomal rearrangement 200 million years ago and retained the overlapping structure ever since (Shintani et al., 1999). The exon overlap of the MINK and CHRNE genes in humans was shown to arise from mutations in the polyA signal of the CHRNE (Dan et al., 2002). A UTR extension of BLZF1 gene has resulted in an overlap with gene C1orf114 in human (Makałowska et al., 2007). On the genome level, Makalowska et al. (2005) examined overlapping genes in vertebrates and showed gene birth, extension of gene ends, or loss of the intergenic sequence between genes also contribute to gene overlaps. They later found repetitive elements play a role in the birth of overlapping genes through transposition and exaptation (Makałowska et al., 2007). Sanna et al. (2008) found 3’UTR evolution is a major contributor to overlap formation between neighboring genes. A study in mosquitos revealed simple sequence repeats are responsible for overlap formation through gene rearrangement (Behura & Severson, 2013). The studies published so far are based on a limited number of cases. More comprehensive studies on

genome level structure of overlapping structure and the conservation of these arrangements are essential to fully understand the significance of gene overlaps and their evolution.

3) Overlapping genes are not only in physical proximity on chromatin but they physically overlap so the antagonizing factor of co-regulation such as transcription interference must be taken into consideration. If the overlapping genes are transcribed from the same strand, this could induce physical constraints, such as transcription collision (Crampton et al., 2006) and mis-splicing of introns (Lee & Chang, 2013) Alternatively, if the genes are transcribed from different strands, then the messages would be complementary to each other leading to alterations in gene expression through RNA degradation via a microRNA pathway (Fire et al., 1998). This transcriptional interference points towards potential negative correlations of expression across tissues or across different stages of the life cycle. A case study by Henikoff et al. (1986) found a nested pair where a pupal cuticle protein was nested within an intron of purine pathway gene and found the former was only expressed for a 3-h window in development while the host gene is a housekeeping gene and thus supporting the interference model. On the genomic level, the picture is more complicated due to the tradeoffs between the positive correlation predicted by co-expression and negative factor of transcriptional interference. For example, co-expression has been observed in *C. elegans* (Chen & Stein, 2006) and in humans (Chen et al., 2019; Soldà et al., 2008), but different overlapping structures show different expression patterns, and the correlation level varies in different overlapping types. On the other hand, the studies focusing on the nested groups generally display less correlated patterns (Assis et al., 2008; Lee & Chang, 2013) or even a negative correlation (Yu et al., 2005).

1.4 Recent progress on overlapping genes since 2015

Since 2015, the cost of sequencing has been drastically reduced and the sequencing quality has increased monumentally, which allows for better annotation data in model as well as

non-model organisms. Long read assembly using PACBIO and Oxford Nanopore technologies has improved assemblies through repetitive regions of the genome. Annotation pipelines have improved as well so that antisense transcripts could be reliably detected and differentiated from sense transcripts.(Wright et al., 2022) Increasing many small ORFs could be captured by incorporating new sequencing technology with proteogenomic methods and CRISPR functional screens, and more potential overlapping genes could be found (Chen et al., 2020). Indeed, a revisit of human overlapping protein-coding genes (Chen et al., 2019) found a significantly higher overlapping percentage 25% compared to 14% in previous studies (Veeramachaneni et al., 2004).

More detailed studies have been done to answer unresolved questions in specific types of overlapping genes. For instance, one study focusing on CDS overlapping in humans (Bukhnikashvili, 2023) found protein overlapping genes tend to: (1) be GC rich due to the lack of stop codons and codon usage bias of amino acid codons, (2) be SNP poor due to selection constraint posed on mutations and (3) have more disordered structure than typical non-overlapping ones. Assis (2016) has updated the evolution model on nested genes versus Assis et al. (2008) using expression data in Drosophila and suggest simultaneous expression of nested pairs are selected against due to transcription interference. Another study Inagaki et al. (2021) has uncovered one of the mechanisms which coordinate the transcription of convergent overlapping pairs on genome level using chromatin profiling in *Arabidopsis*. They found that a histone demethylase FLD and its associating factor can remove the histone H3HK4me mark within the convergent overlapping region and thus downregulate their expression when the antisense transcriptions are high. Moreover, synthetic approaches have been utilized in virus and prokaryotes to study the consequence of overlapping sequences (Wright et al., 2022) and the conservation of overlapping sequences is taken advantage of by bioengineers to prolong the stability of genetic circuit in bacteria (Chlebek et al., 2023).

1.5 The phylogeny of the 12 *Drosophila* species.

D. melanogaster is one of the best studies model systems from the inception of genetics and evolutionary research. The *D. melanogaster* genome is well-annotated with over 13,000 genes. And it is also unique because the genome is compact and rearrangement-rich (Bhutkar et al., 2008), making it suitable as a reference genome to studying overlapping genes.

There are eleven other related *Drosophila* species that were selected representatives from the genus *Drosophila* phylogeny. They have been sequenced, assembled and annotated in the 2000s and extensive research on molecular evolution has been done since based on the comparative features among the 12 species (Consortium et al., 2007; Richards et al., 2005; Schaeffer et al., 2008). The set of species diverged around 43 million years ago (Kumar et al., 2017) and form two subgenera *Drosophila* and *Sophophora* and further evolved into multiple species groups at various time points (Figure **1-5-1**). *D. pseudoobscura* and *D. persimilis* are from the obscura group. *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae* are more closely related to *D. melanogaster* which form the melanogaster group. *D. willistoni*, *D. grimshawi, D. virilis*, *D. mojavensis* are more distantly related with *D. melanogaster* and the latter three from subgenus *Drosophila.* These species are rather diversified in their morphology and habitats, but their genomes are relatively conserved in many aspects such as scale of genome size and number of genes (Consortium et al., 2007). Particularly, their genomes are distributed among five chromosomes arms and a small dot chromosome. Each chromosome arm corresponds to a Muller element: element A corresponds to the X chromosome, elements B to E corresponds to the four autosomal arms and element F corresponds to the small dot chromosome (Schaeffer 2018). Muller elements provide a method for comparing homologous chromosome arms (Table **1-5-1**). They are homologous across the 12 species and yet there are extensive fusion of chromosomal arms and inversion events on a macro scale and plenty of gene rearrangement events within and between each Muller element on a micro scale (Bhutkar et al., 2008). Overall,

the availability of genome sequences for 12 species from the *Drosophila* and *Sophophora* subgenera provide a suitable time depth to track evolution of overlapping genes and allow us to use comparative genomics to investigate the evolution of overlapping genes.

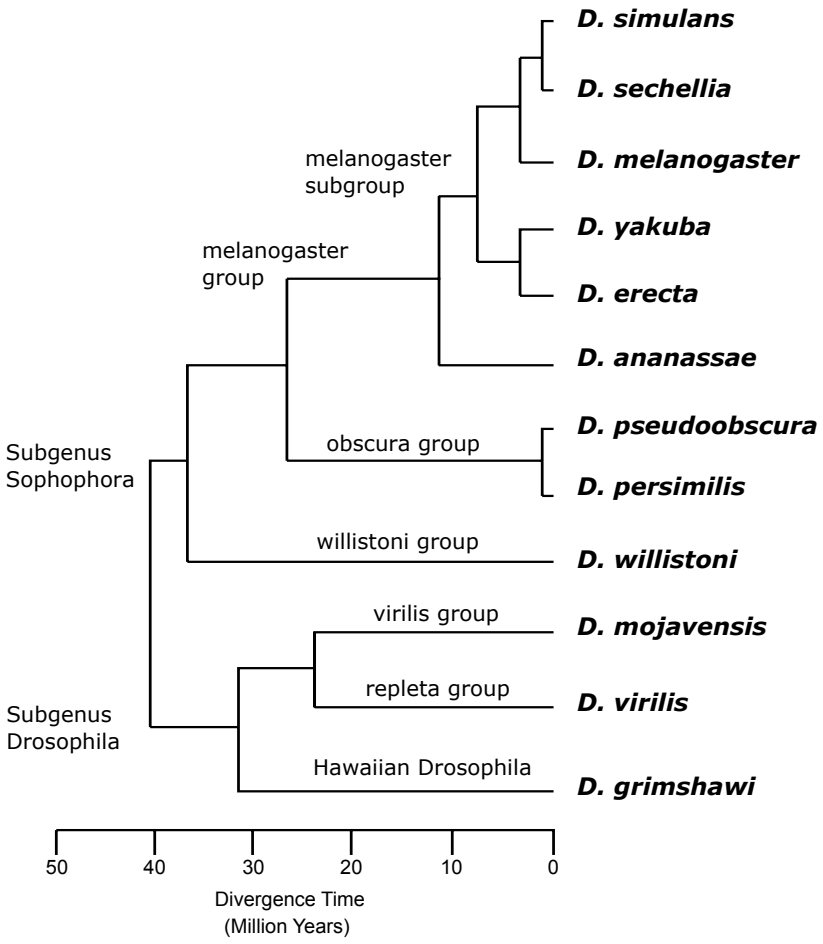


Figure **1-5-1**: The phylogeny of the 12 fly species (Source: adapted from Schaeffer et al. 2008).

Table **1-5-1**: Muller elements equivalence table (Source: adapted from Schaeffer et al. 2008).

	Muller A	**Muller B**	**Muller C**	**Muller D**	**Muller E**	**Muller F**
D. simulans	Chr X	Chr 2L	Chr 2R	Chr 3L	Chr 3R	Chr 4
D. sechellia						
D. yakuba						
D. erecta	Chr X	Chr 2L, 2R	Chr 2L, 2R	Chr 3L	Chr 3R	Chr 4
D. ananassae	Chr XL, XR	Chr 3R	Chr 3L	Chr 2R	Chr 2L	Chr 4L, 4R
D. pseudoobscura	Chr XL, XR	Chr 4	Chr 3	Chr XR	Chr 2	Chr 5
D. persimilis						
D. willistoni	Chr XL	Chr 2R	Chr 2L	Chr XR	Chr 3	Chr 3
D. virilis	Chr X	Chr 4	Chr 5	Chr 3	Chr 2	Chr 6
D. mojavensis	Chr X	Chr 3	Chr 5	Chr 4	Chr 2	Chr 6
D. grimshawi	Chr X	Chr 3	Chr 2	Chr 5	Chr 4	Chr 6

Note: Chr stands for Chromosome

1.6 Summary of work done.

In sum, overlapping genes are a universal and interesting phenomenon and whether they are a neutral or selected phenomenon in eukaryotic genomes is not clear. In this study we approached this overarching question on the gene structure level from the following three aspects: 1) how common are overlapping genes in the *Drosophila* and whether overlapping gene structure are conserved, 2) how does overlapping structure evolve during speciation; 3) how do overlapping genes behave under positive effect of transcriptional co-regulation and negative effect of transcription interference. We used *D. melanogaster* as our reference genome to study the evolution and expression pattern of overlapping genes in the 12 fly species. The second chapter focuses on the general scheme design and methods applied. The third chapter displays all the results we found. The fourth chapter gives our interpretation of the work we have done. Overall, we found over 30 % of the genes in *D. melanogaster* overlap. They are generally more conserved than non-overlapping genes and the degree of conservation correlates with the configuration of the overlap, namely relative orientation and proportion of overlap. We found that

the level of conservation of overlapping genes relates to their transcription regulation and general function in development and are not due to functional interaction between partners. We demonstrated different overlapping configurations also show different evolutionary patterns. We showed other than physical cluster effect, overlapping could be another factor, which have distinctive expression and evolutionary pattern than non-overlapping gene clusters.

This work was mostly done in the period of 2010-2015. The writing and interpretation of results are done in the fall of 2024 incorporating the newest advances in this field. This project was halted in between due to personal reasons. Given the limitation of extended time of this thesis, all the data used are based on the data available at that time. Nonetheless, the investigation done based on old data remains relevant as methods used are robust and no systematic study has been published so far on the overlapping genes evolution and expression across the 12 fly species.

Chapter 2

Materials & Methods

2.1 Experimental design.

First, we used *D. melanogaster* as our reference genome to identify and classify the overlapping genes. We also set up two control groups. The first is the background control set which is randomly selected from all non-overlapping genes in *D. melanogaster* genome. A second control set is non-overlapping neighbor pairs selected from the genomic background to discriminate the effect of physical overlapping versus physical clustering effect. (Section 2.2)

Based on the gene identifiers in *D. melanogaster* we then searched the orthologs of overlapping genes in the 11 non-*D. melanogaster* species. During this process we found cases of putative artifacts due to annotation issues in the 11 species which could confound ortholog calls in non- *D. melanogaster* species upon manual examination. Therefore, we set up additional pipelines to detect putative artifacts (Section 2.3 & 2.4). Based on the orthologs calls we then reconstructed the ancestral states of overlapping genes in the 12-fly phylogeny (Section 2.5).

Finally, we used strand specific RNA-Seq data to investigate expression patterns of overlapping genes in *D. melanogaster* (Section 2.6).

2.2 Classification of overlapping genes.

The genome of *Drosophila melanogaster* is the best-annotated genome of all the 12 fly species whose genomes have been sequenced to date. Thus, we used *D. melanogaster* as our reference genome to annotate the overlaps among genes. The genome of *D. melanogaster* (release 5.22) was retrieved from Flybase (Tweedie et al., 2009). The coordinates of all protein coding genes served as the source of information to score whether two genes overlapped. We defined

two genes as overlapping if any of the sequence between their transcriptional start and stop sites were coincident (greater or equal to 1bp). For genes with alternative splicing, we used the coordinates of the longest transcript. Detailed classification of overlapping subgroups was based on comparisons of overlapping gene pair coordinates and relative strand orientations. Namely, for overlap to be complete, one gene's coordinates should be within the range of the other, if not the overlapping pair is defined to be partial. For the relative orientation of overlap, we marked the 5' to 3' chromosome strand as the positive sign and the 3' to 5' chromosomal strand as the negative sign. Same strand overlap requires both genes to have the same signs, and vice versa.

In addition, we set up control groups by using all non-overlapping genes in *D. melanogaster* which are 500 bp apart to eliminate the confounding factor of promoter overlaps according to reported length of core promoter (Chen & Stein, 2006) All non-overlapping genes were then randomly shuffled to generate the non-overlapping random pair set. All random pairs were pooled to generate the non-overlapping random group. The non-overlapping neighbor pair set was generated by choosing non-overlapping genes that are adjacent to each other. All non-overlapping neighbor pairs were pooled to generate the non-overlapping neighbor group.

2.3 Gene annotation artifacts detection using overall pairwise alignments against *D. melanogaster.*

We evaluated two scenarios where annotation artifacts might lead to inaccurate inference of orthologs in the non-*D. melanogaster* species (Figure **2-3-1**). First, in the overlapping pairs, exons of one overlapping gene of the pair might be annotated as a separate gene resulting in an extra gene model between overlapping orthologs in non-*D. melanogaster* species, which introduce artifacts in gene order difference calculation. Second, in the parent-embedded parallel case, the embedded ortholog could be annotated as one exon of the parent genes. Analysis was

performed for all the non-*D. melanogaster* species. We then counted the total number of putative artifacts cases and calculated the percentage of annotation artifacts for each species.

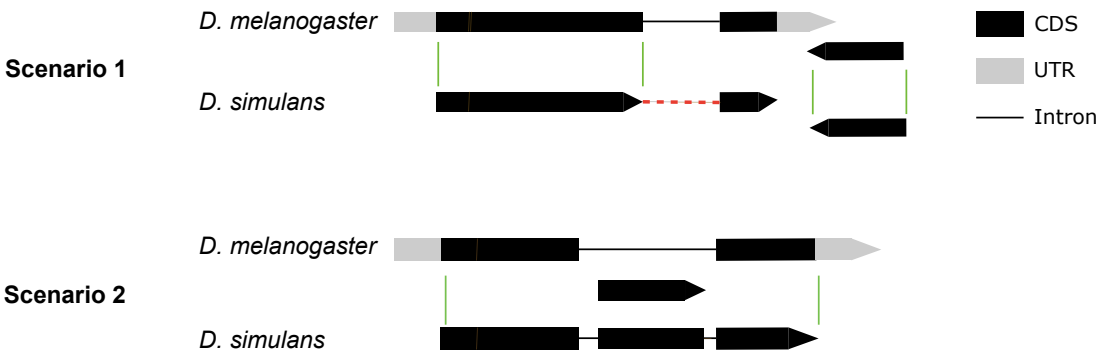


Figure **2-3-1**: Illustrations of putative artifacts. Scenario1: the orthologous sequences of the second exon of the upstream gene of the overlapping pair in *D. melanogaster* is annotated as a separate gene in *D. simulans*. Scenario 2: the embedded ortholog is annotated as an exon of the parent gene in *D. simulans*.

If we take one of the non-*D. melanogaster* species-*D. simulans* as an example. The most recent overall pairwise alignments against *D. melanogaster* for *D. simulans* were retrieved from UCSC website (https://hgdownload.soe.ucsc.edu/downloads.html#fruitfly, dm3, Apr. 2006, BDGP Release 5). In the first case, we used the CDS coordinates of the *D. melanogaster*'s overlapping gene with the same direction as the interval gene as query and retrieved the alignment segments and looked up the coordinates in *D. simulans*. If *D. simulans*' corresponding segment contains both orthologs of the overlapping gene and the interval gene, we deemed it as a putative artifact. Otherwise, if *D. simulans*' corresponding segment only contains the canonical ortholog, the interval gene was deemed as a true *D. simulans*' species-specific gene.

For the second case, we used the CDS coordinates of the parent gene in *D. melanogaster* as the query and retrieved the alignment segment and looked up the coordinates in *D. simulans*. Next, we used the CDS coordinates of the embedded gene in *D. melanogaster* and retrieved the alignment segment and looked up the coordinates in *D. simulans*. Then we compared the two coordinates sets in *D. simulans*. If the embedded gene's corresponding coordinates coincident with exon regions of *D. simulans*' parent gene, then it was deemed as a putative artifact. Otherwise, it was deemed as a true missing embedded gene.

2.4 Estimates of conservation of overlapping genes.

We retrieved the gene annotation data of non-*D. melanogaster* species from Flybase (Tweedie et al., 2009). The version numbers are as follows: *D. simulans* (Release 1.3), *D. sechellia* (Release 1.2), *D. erecta* (Release 1.3), *D. yakuba* (Release 1.3) *D. ananassae* (Release 1.3), *D .pseudoobscura* (Release 1.3), *D. persimilis* (Release 1.3), *D. willistoni* (Release 1.3), *D. virilis* (Release 1.3), *D. mojavensis*(Release 1.3), *D. grimshawi* (Release 1.3), we first indexed the genes with an ID number beginning with Muller A and ending with Muller F, and then calculated gene order differences across the phylogeny. The Muller element nomenclature is used as a system to compare genes from syntenic chromosomes (Schaeffer, 2018). For *D. melanogaster*, we used cDNA coordinates start sites as the ordering criteria. Pairwise comparisons between *D. melanogaster* and non-*D. melanogaster* species were done and order differences in the latter were calculated. If gene order difference of the ortholog pair in non-*D. melanogaster* species is one meaning both ortholog pairs are present and are adjacent, they were deemed conserved. Any gene order difference other than one that reflects the gene level evolutionary events such as gene gain, loss and rearrangements and thus were deemed non-conserved across species.

2.5 Ancestral state reconstruction based on the maximum parsimony assumption.

We used the gene order system (Figure **2-5-1**) to investigate the evolutionary history of overlapping pairs. We indexed all protein coding genes in each species consecutively based on their transcripts coordinates. The number associated with each gene is called gene order and orthologous mappings across different species were established using their gene orders. Using this system, we then inferred the evolutionary events based on gene order difference between *D. melanogaster* and non-*D. melanogaster* species. Figure **2-5-2** shows the gene order differences of five orthologous pairs between *D. melanogaster* and non-*D. melanogaster* species revealed an overlapping structure breakage event by gene rearrangement. For each overlapping pair, we then defined different states in each species based on the gene order differences of *D. melanogaster* and their orthologs in non-*D. melanogaster* species (Table **2-5-1**): two orthologs in non-*D. melanogaster* species could overlap as in *D. melanogaster* (state 1); the upstream ortholog could be missing (state 2); the downstream ortholog could be missing (state 3); both orthologs could be missing (state 4); both orthologs could be found at different locations in the non-*D. melanogaster* genome (state 5). We then assigned states to all the overlapping orthologs in all the fly species. Under the maximum parsimony assumption, except for state 1, which indicates overlap conservation across all 12 species, each other state indicates one of several evolutionary events on gene level resulting from an overlapping structure change in the non-*D. melanogaster* species: state 2,3 and 4 correspond to gene gains and losses events; state 5 corresponds to rearrangement events.

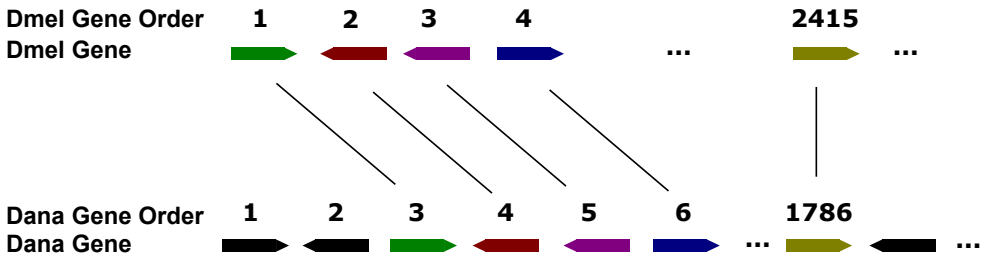


Figure **2-5-1**: The gene order system used to investigate the evolutionary history of overlapping pairs. The colored arrows denote the gene models in the species. All genes are indexed consecutively based on their transcripts coordinates. The number associated with each gene is called gene order. The lines here denote the orthologous mapping between two species.

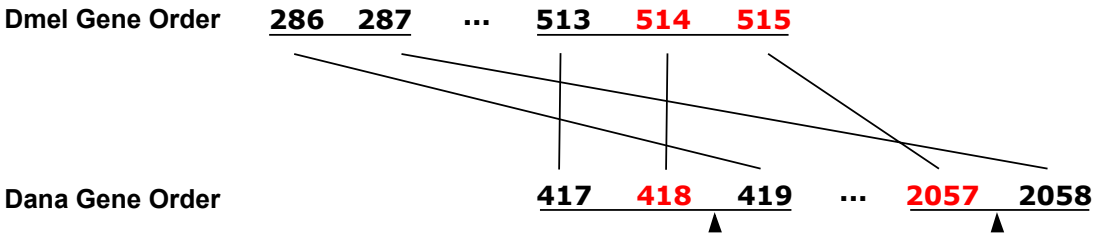


Figure **2-5-2**: A gene inversion event results in the formation of an overlapping pair in *D. melanogaster*. The numbers denote gene orders. Black numbers denote non-overlapping genes and red numbers denote overlapping genes in *D. melanogaster* and their orthologs in *D. ananassae*. Lines in between two species indicate orthologous mapping. Black arrows indicate breakpoints on the chromosome.

Table **2-5-1**: A hypothetical example of states inference.

Dmel Gene 1 Order	100	100	100	100	100
Dmel Gene 2 Order	101	101	101	101	101
Dmel States	**State 1**	**State 1**	**State 1**	**State 1**	**State 1**
Dana Ortholog 1 Order	500	NA*	500	NA	500
Dana Ortholog 2 Order	501	501	NA	NA	1000
Dana States	**State 1**	**State 2**	**State 3**	**State 4**	**State 5**

Evolutionary Events Responsible for State Change	Conserved	Gene Loss/Gain	Gene Loss/Gain	Gene Loss/Gain	Gene Rearrangement

*NA: no ortholog detected.

We used the maximum parsimony principle to reconstruct ancestral states of overlapping pairs along the evolution of the 12 species. We used the species tree topology from Flybase (Tweedie et al., 2009) and used Sankoff's algorithm (Sankoff, 1975) to assign the ancestral states to each internal node for each overlapping pair. Transition rates between different states were assumed to be equal. We then tracked the total number of gene gains, losses and rearrangement events for each branch on the 12 species phylogeny. All pairs with ambiguous ancestor states were eliminated and the total number of evolutionary events were normalized to get mean events per overlapping pair for each branch and for each overlapping type. Because gene level evolutionary events occurring on each branch are very rare, we assumed a Poisson distribution to calculate the 95% confidence of event rate and marked with the error bars in the phylogeny.

2.6 Gene expression of overlapping genes.

We next investigated the expression pattern of overlapping genes in *D. melanogaster*. To avoid ambiguous read mapping due to the sequence overlaps, we specifically chose the strand specific RNA sequence dataset, which has been demonstrated to be suitable for studying complementary transcripts (Passalacqua et al., 2012; Siegel et al., 2014). The RNA-Seq data (SRA accession number: SRA009364) used was from (Graveley et al., 2011) and was generated by taking 12 total RNA samples from different embryonic stages of *D. melanogaster* and then were sequenced with the strand specific technique. The raw sequencing data were then converted to FPKM values. The FPKM values were log transformed by added one pseudo read to each raw

FPKM value for each gene to allow log2 transformation of the data. The value after log2 transformation was treated as the expression value for the corresponding gene.

Chapter 3

Results

3.1 Overlapping genes overview.

In total, we identified 4,620 (33.91%) overlapping genes out of the 13,626 protein-coding genes in release 5.22 of *D. melanogaster*. They form 1,958 overlapping gene clusters with number of genes within cluster ranging from two to 15 (Figure **3-1-1)**. To find how these overlapping genes are distributed across the genome, we calculated the overlapping gene density with a resolution of 1Mb bins and plotted this against the background whole genome gene density measured with the same resolution. We found that overlapping gene density is strongly positively correlated with the background gene density (Figure **3-1-2,** Pearson's correlation= 0.89, p-value < 2.2e-16). This result indicates that overlapping genes are distributed homogenously across genome. We further filtered the overlapping gene clusters to only keep two-gene overlapping pairs because two-gene overlaps represent the majority of the clusters (79.32%). We eliminated complex gene clusters to remove gene number as a confounding variable to simplify our analysis. Furthermore, among all overlapping genes, we detected 53 CDS overlaps in 40 overlapping pairs (Table **S1**) and found that CDS overlaps are generally short (The longest outlier turns out to be a lnc RNA gene in the most recent release upon manual examination) and there is no strong bias in the distribution of relative orientation and codon frameshift phases. (Figure **3-1-3**). They were eliminated from the following analysis due to their unique sequence context and small sample sizes.

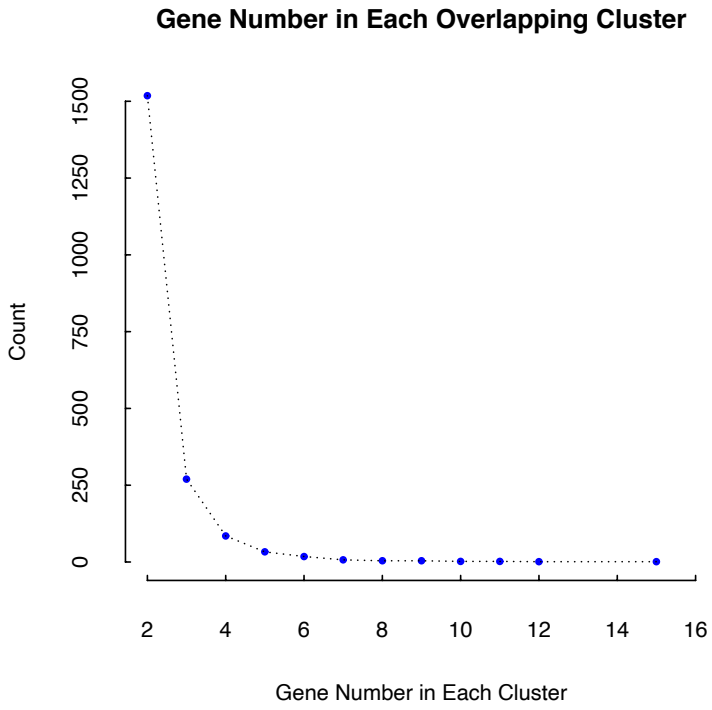


Figure 3-1-1: Distribution of gene number in overlapping clusters. Gene numbers involved in each overlapping cluster are illustrated in this figure. Gene numbers range from 2 to 15. Most clusters contain 2 genes.

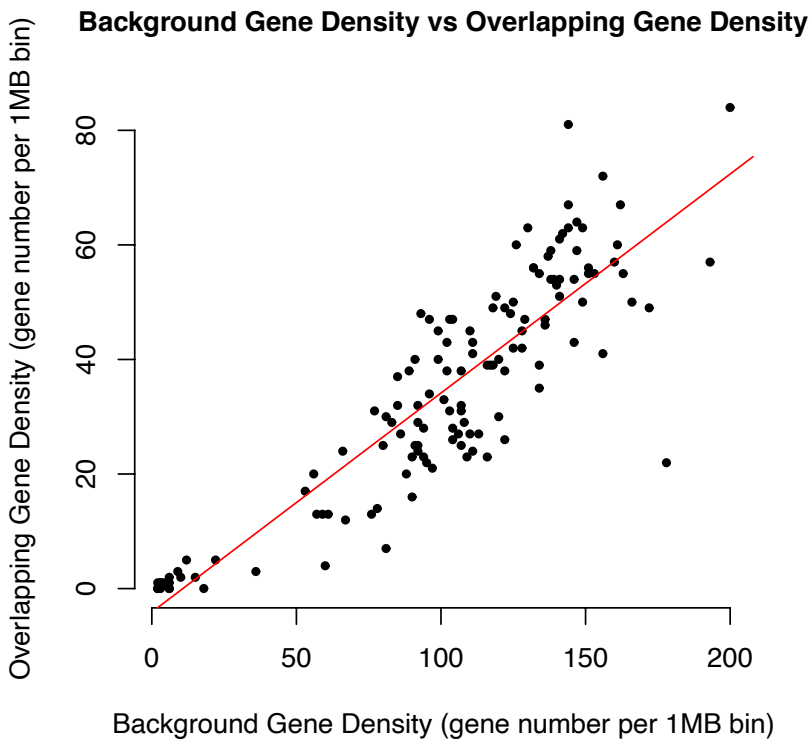


Figure **3-1-2**: Overlapping gene density across genome vs. background gene density. The X-axis indicates the background gene density, and the Y-axis indicates the overlapping gene density. Gene density is defined as the number of genes found per 1Mb bin of genomic sequences. Red line indicates the linear regression line.

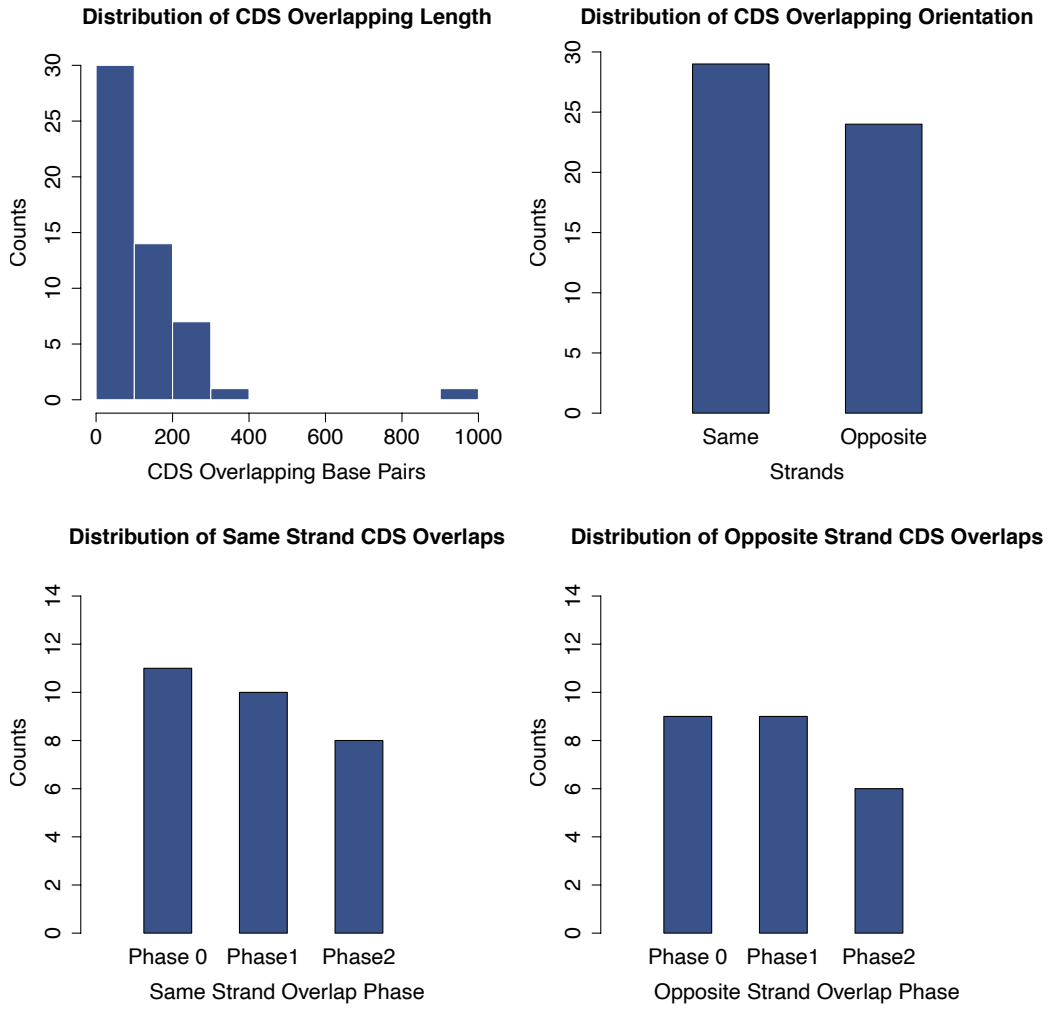


Figure **3-1-3:** Distribution of protein overlapping genes. Top left: distribution of CDS overlapping length. The X-axis indicates the overlapping length, and the Y-axis shows the frequency; Top right: distribution of CDS overlaps in terms of relative orientation. The X-axis indicates whether CDS overlaps are found on the same or opposite strands and the Y-axis shows the frequency; Bottom left: distribution of CDS overlaps on the same strand. The X-axis indicates the codon frame shift phases of CDS overlaps and the Y-axis shows the frequency; Bottom right: distribution of CDS overlaps on the opposite strands. The X-axis indicates the codon frame shift phases of CDS overlaps and the Y-axis shows the frequency.

Two major factors contribute to variation in overlapping configurations, the length of the overlap segment and the relative orientation of the overlapping pair. A partial overlap occurs

when only a portion of the two genes shares sequence. A complete overlap occurs when one gene is fully encompassed within the other gene. Orientation matters because overlapping pairs could reside either on the same strand or opposite strands. Based on different combinations of these two variables we classified overlapping pairs into five different overlapping configurations (Figure **3-1-4**): 1) convergent overlapping pair where the overlap is partial and 3' end to 3'end; 2) divergent overlapping pair where the overlap is partial and 5'end to 5'end; 3) parallel overlapping pair where the overlap is partial and 3'end to 5' end; 4) parent-embedded anti-parallel pair where the overlap is complete and on opposite strands; 5) parent-embedded parallel pair where the overlap is complete and on the same strand. In total, we identified 787 convergent overlapping pairs, 25 divergent overlapping pairs, 40 parallel overlapping pairs, 207 parent-embedded anti-parallel pair, and 112 parent-embedded parallel pair. This bias of numbers for different overlapping configurations are illustrated in Figure **3-1-5,** which shows proportions and direction of each gene that is overlapped with its partner. Overlapping pairs form three clusters: the center with a diameter less than 0.5 and two parallel lines at the top and bottom, which means genes either overlap of a small proportion of themselves, such as UTR regions of 3' or 5'end or with a complete overlap, such as the intron of the longer pair, but not in between. We also found that there is a significant bias towards opposite strands overlapping versus same strand overlaps (1019:152, p-value <2.2e-16).

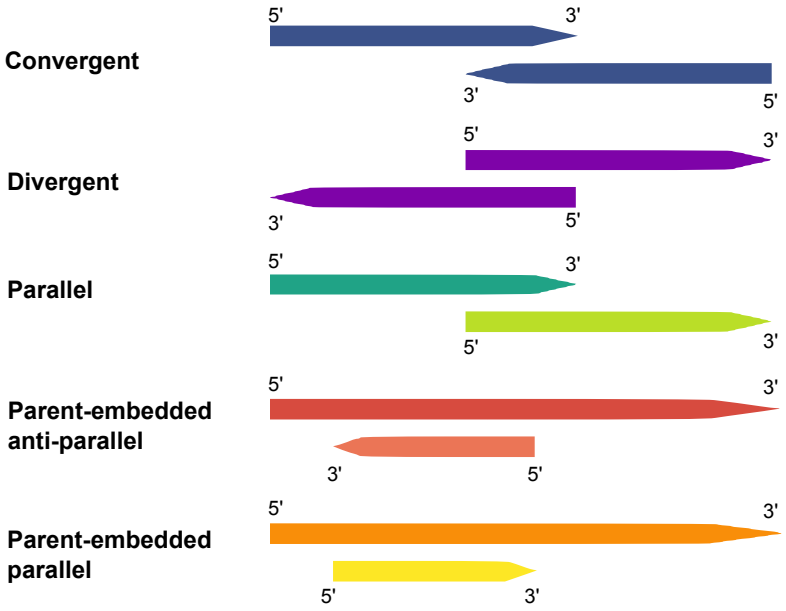


Figure **3-1-4**: Classification schemes of two gene overlaps. Different colors represent different overlapping types. The direction of the arrows indicates the direction of 5' to 3'. This figure shows schematically five different overlapping configurations: 1) convergent overlapping structure; 2) divergent overlapping structure; 3) parallel overlapping structure; 4) parent-embedded anti-parallel structure; 5) parent-embedded parallel structure.

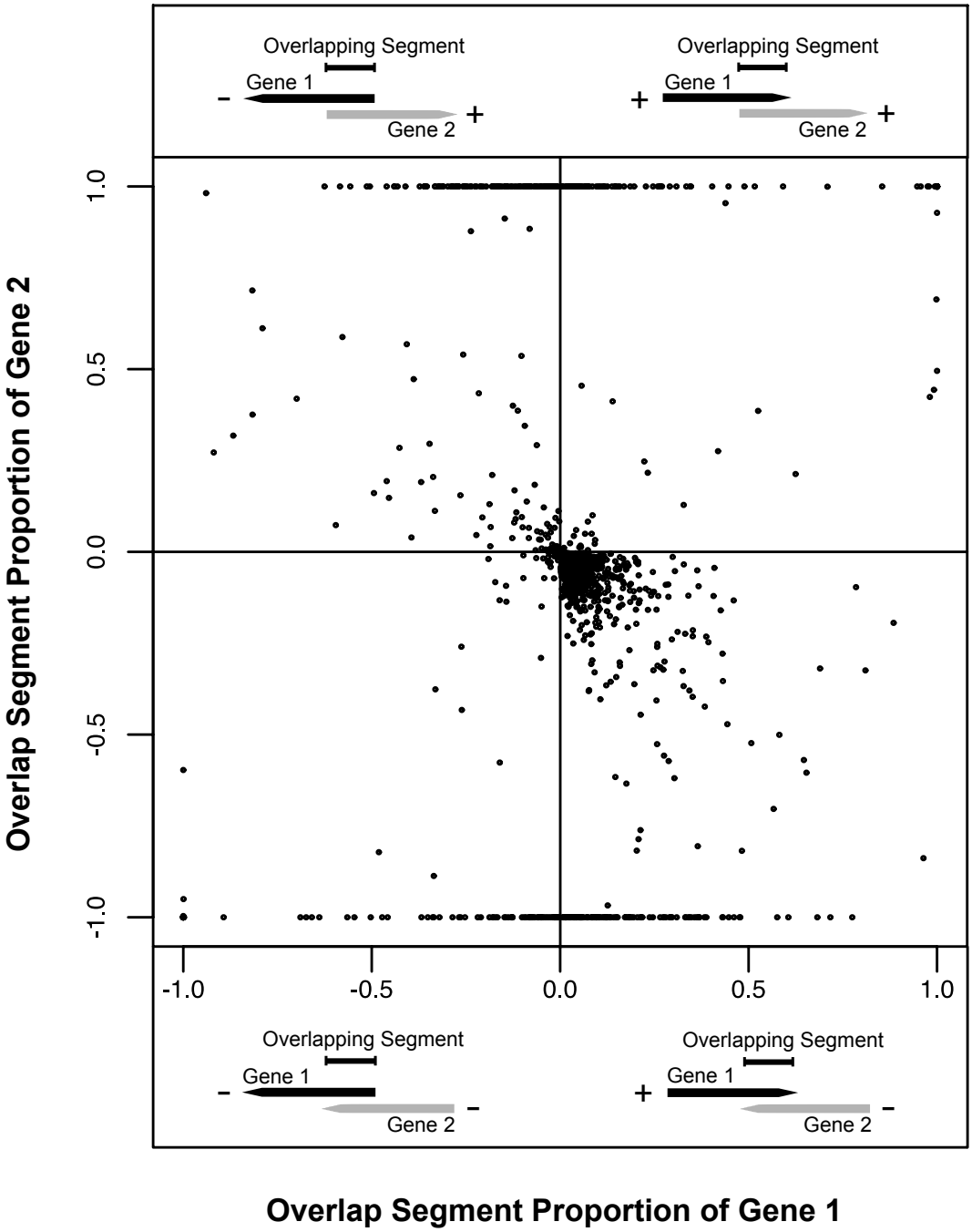


Figure **3-1-5**: Two gene overlapping segments and orientation. This figure shows oriented proportions of each gene that is overlapped with its partner. Gene one is the left gene of the two-overlapping pair. Gene two is the right gene of the two-overlapping pair. Positive sign indicates 5'-3' direction and negative sign indicates 3'-5' direction.

3.2 Conservation degree estimation of overlapping structure

Overlapping genes are susceptible to annotation artifacts due to sharing sequence, especially for the non-*D. melanogaster* species whose annotation at the time of this analysis was in the primary version without later manual curation from experimental and long-read sequencing data. Therefore, we first evaluated the annotation quality of overlapping orthologs in the non- *D. melanogaster* to distinguish evolutionary events occurred in the other species from gene model annotation artifacts in the other species. For example, in the overlapping cases, exons of the overlapping genes were annotated as separate genes. In the parent-embedded parallel case, the embedded ortholog could be annotated as one exon of the parent genes (see methods).

In the first case, we did not detect putative artifacts in parallel overlapping and divergent overlapping pairs. For convergent overlapping pairs, we detected 32 pairs out of 787 pairs (4.07%) might have putative artifacts in one or more non-*D. melanogaster*. Putative artifacts might be a more serious problem for parent-embedded parallel pairs. We detected 43 pairs out of 112 pairs (38.39%) with putative artifacts in one or more additional species. (Table **3-2-1**). We discarded these putative artifacts in the 12 fly species evolutionary analysis.

Table **3-2-1**: Putative artifacts detected in convergent overlapping pairs and parent-embedded parallel pairs.

Species	**Putative Artifacts**	
	Convergent Overlapping (Percentage)	**Parent-embedded Parallel (Percentage)**
D. simulans	13 (1.65%)	10 (8.93%)
D. sechellia	0 (0%)	0 (0%)
D. yakuba	7 (0.89%)	17 (15.18%)
D. erecta	6 (0.76%)	14 (12.50%)
D. ananassae	3 (0.38%)	10 (8.93%)
D. pseudoobscura	7 (0.89%)	13 (11.61%)
D. persimilis	0 (0%)	0 (0%)
D. willistoni	0 (0%)	0 (0%)
D. virilis	8 (1.02%)	8 (7.14%)

D. mojavensis	7 (0.89%)	11 (9.82%)
D. grimshawi	3 (0.38%)	10 (8.93%)
Total	**32 (4.07%)**	**43 (38.39%)**

Because the cis regulatory regions such as 3'UTRx and 5'UTRx or introns could have important functions to maintain expression levels of both genes in an overlapping pair (Martins et al. 2011) we hypothesized that genes whose regulatory sequences overlap will tend to be more conserved in their gene order than non-overlapping genes.

Conservation degree was calculated according to the description in the Methods in all non-*D. melanogaster* species. In total, we found 72.32% convergent overlapping pairs, 28.00% divergent overlapping pairs, 57.50% parallel overlapping pairs, 48.79%parent-embedded anti-parallel pairs, 23.19% parent-embedded parallel pairs are conserved, compared to 34.05% non-overlapping neighbor genes conserved across all species.(Table **3-2-2**) These results indicate that the degree of conservation for different overlapping types varied, with convergent overlapping pairs being significantly more conserved than all other groups and the parent-embedded being the least conserved of all. Interestingly, full overlaps, i.e., parent-embedded pairs, regardless of orientation, are significantly less conserved than convergent overlapping pairs, and are not significantly different from non-overlapping neighbor controls. We also investigated the conservation level of overlapping pairs among the 12 fly species. (Figure **3-2-1**). We found the conservation degree of overlapping pairs was generally negatively correlated with divergence time, except for *D. sechellia* and *D. simulans*.

Table **3-2-2**: Conservation ratio for different types of overlapping pairs.

Overlapping Types	**Total Number of Conserved Pairs**	**Total Number of Pairs**	**Conservation Percentage**	**95% CI of Conservation Percentage**
Convergent Overlapping	546	755	72.32%	72.32±3.19%
Divergent Overlapping	7	25	28.00%	28.00±17.60%
Parallel Overlapping	23	40	57.50%	57.50±15.32%
Parent-embedded Antiparallel	101	207	48.79%	48.79±6.81%
Parent-embedded Parallel	16	69	23.19%	23.19±9.96%
Non-overlapping Neighbor	584	1715	34.05%	34.05±2.24%

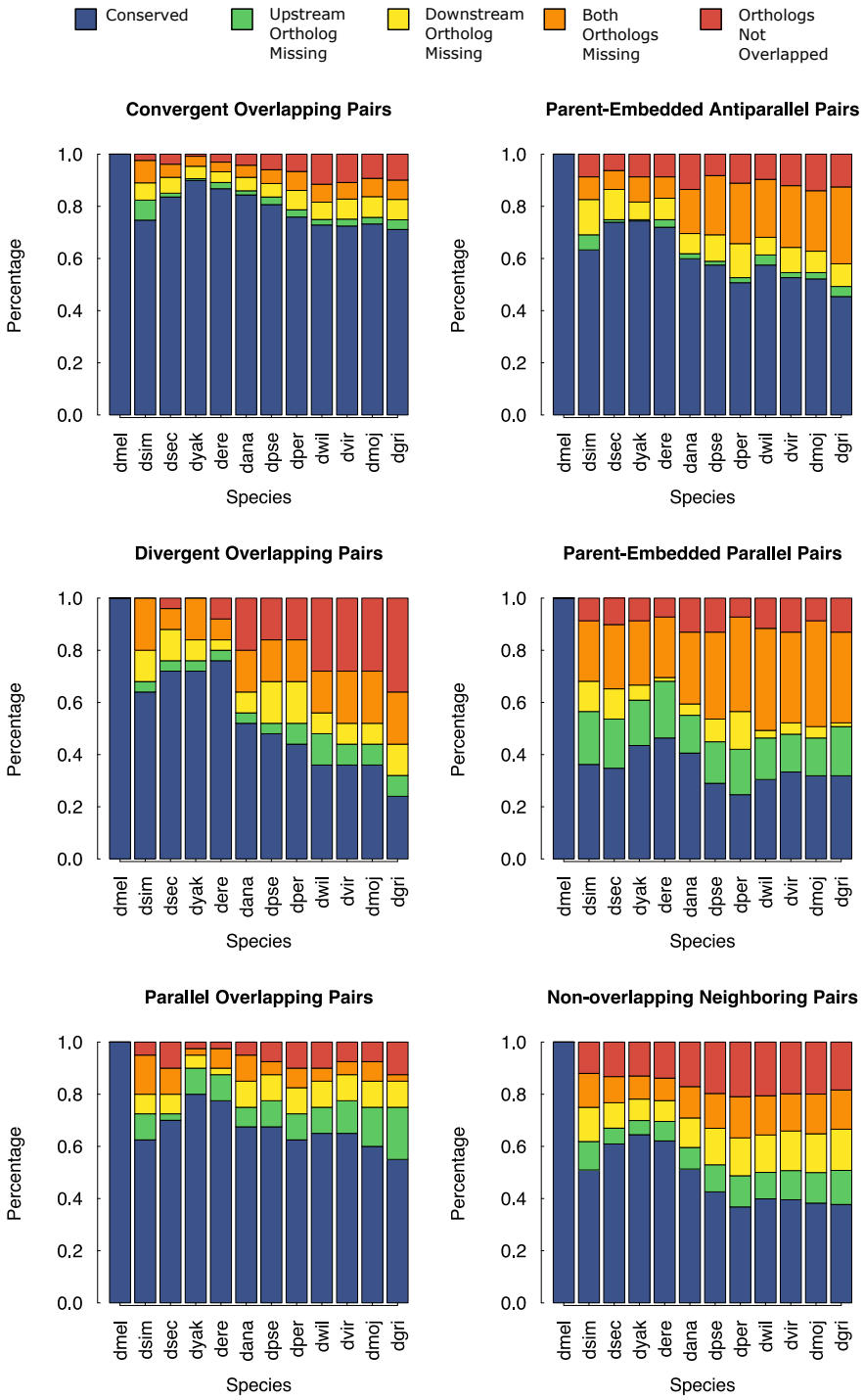


Figure **3-2-1**: Conservation degrees across the 12 fly species. X axis indicates different species and Y axis indicates the percentage of conserved pairs and different scenarios of non-conserved pairs.

3.3 The origin and evolutionary history of overlapping genes.

We next investigated the origin and evolution of overlapping genes. Several mechanisms have been proposed including gene gain, gene loss and gene rearrangement such as transposition and inversion (Calvete et al., 2012; Makałowska et al., 2007; Wright et al., 2022). To model such discrete gene level change events along the evolution of the 12 fly species, we used a gene order index system which allows us to do numerical inferences. We retrieved gene orders of the ortholog pairs of *D. melanogaster*'s overlapping pairs in the other 11 fly species and estimated the state transitions on each branch accordingly (see methods). Under maximum parsimony principle we then calculated the average number of evolutionary events (gene gains, losses, and gene rearrangements) occurred on each branch throughout the phylogeny of the twelve species (Figure **3-3-1)**.

Overall, we found overlapping structure changes the most on the branches leading to *D. melanogaster.* Particularly, the *D. melanogaster* branch is enriched with overlapping pairs formation through *de novo* gene gain i.e., new genes are born near pre-existing or within pre-exiting genes and results in new overlaps. Gene rearrangements play a lesser role in the formation of overlapping structure in general. Recurrences of overlapping pairs in other species (red and blue boxes in non-*D. melanogaster* branch) are rare suggesting convergent evolution of overlapping structure are uncommon. On the other hand, we found new overlapping pairs once formed in ancestor species could then go through multiple losses along the evolution in different lineages, indicating these overlapping structures are not static. Compared to overlapping structure formation, the rate of losses is homogeneous during speciation of non-*D. melanogaster* species, except for *D. simulans*. In addition, species show different patterns of overlapping gene evolution. Gene losses (green boxes) is the major cause of overlapping structure loss in *D.*

simulans and *D. persimilis*, while gene rearrangements (yellow boxes) caused breakages are the predominating cases in *D. willistoni*. Moreover, the evolutionary history of different overlapping structures also varies. Parent-embedded parallel pairs evolved almost exclusively in the *D. melanogaster* branch through gene gain. Divergent overlapping pairs have very few gene loss events and gene rearrangements play a predominating role in the formation of overlapping structure. In contrast, gene loss/gain is quite common for the formation and breakage of parallel overlapping structure. (**Figure 3-3-2**)

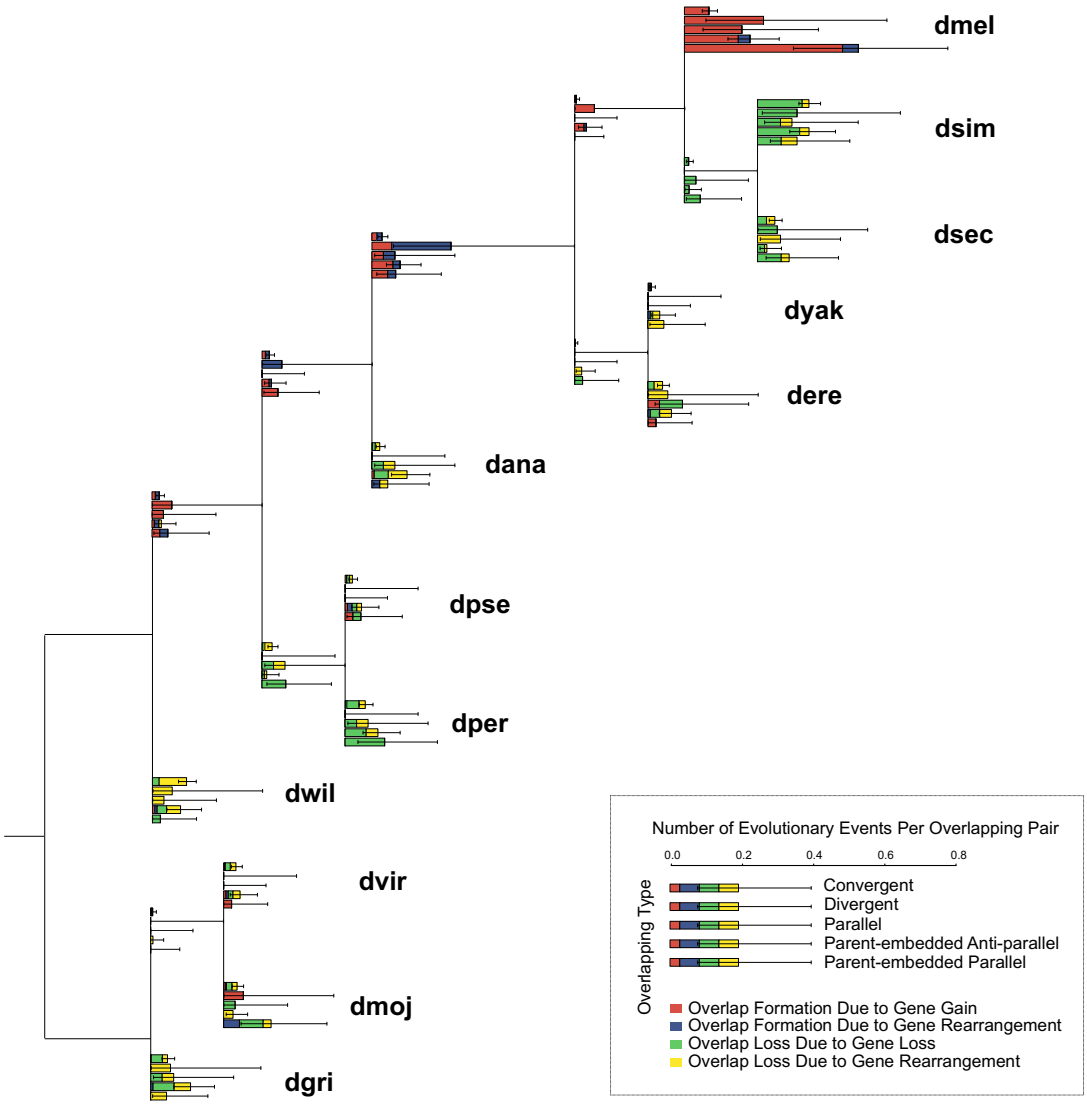


Figure **3-3-1**: Evolution of overlapping pairs in the 12 fly species. Legend is shown on the right bottom corner. The horizontal length of each box indicates mean evolutionary events per overlapping pair. Different colors represent different evolutionary events. Error bars indicate 95% Poisson Confidence Interval.

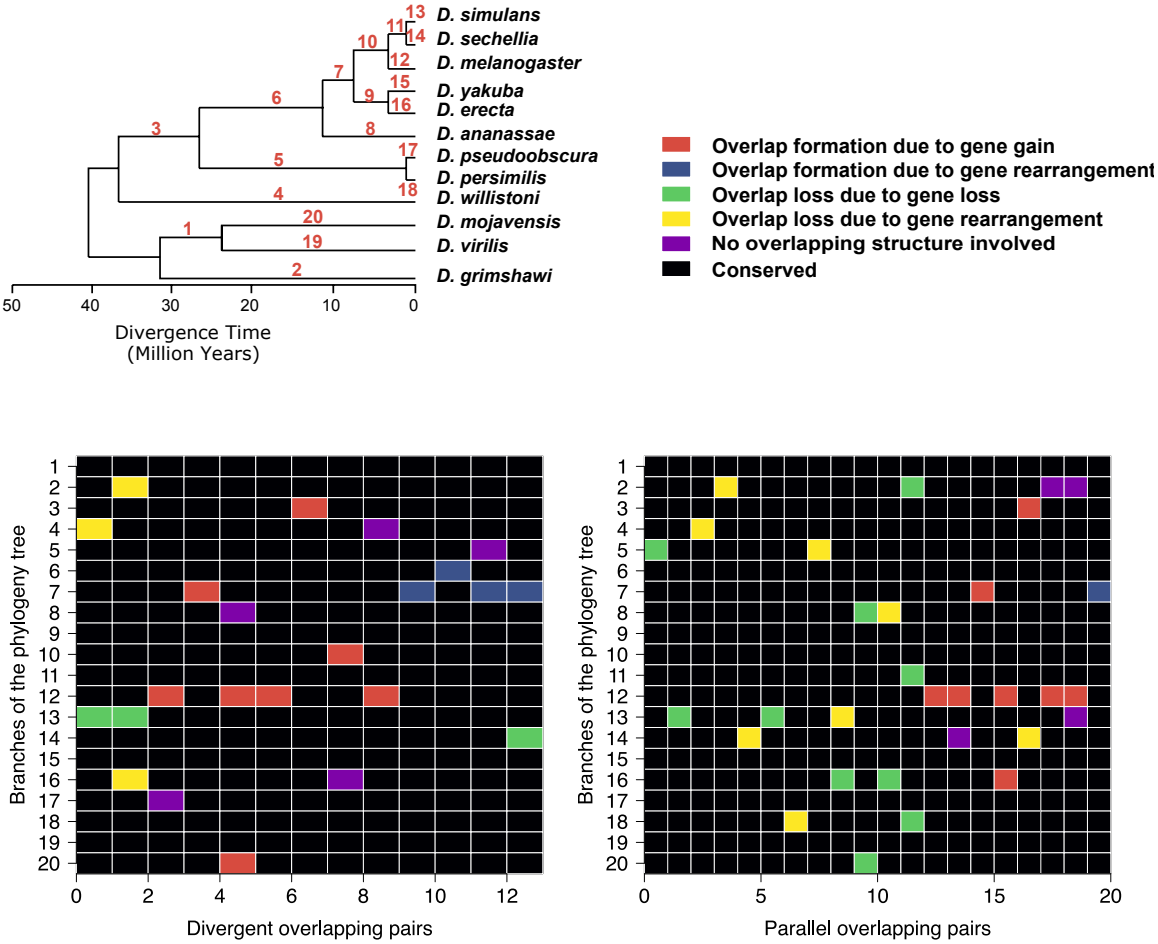


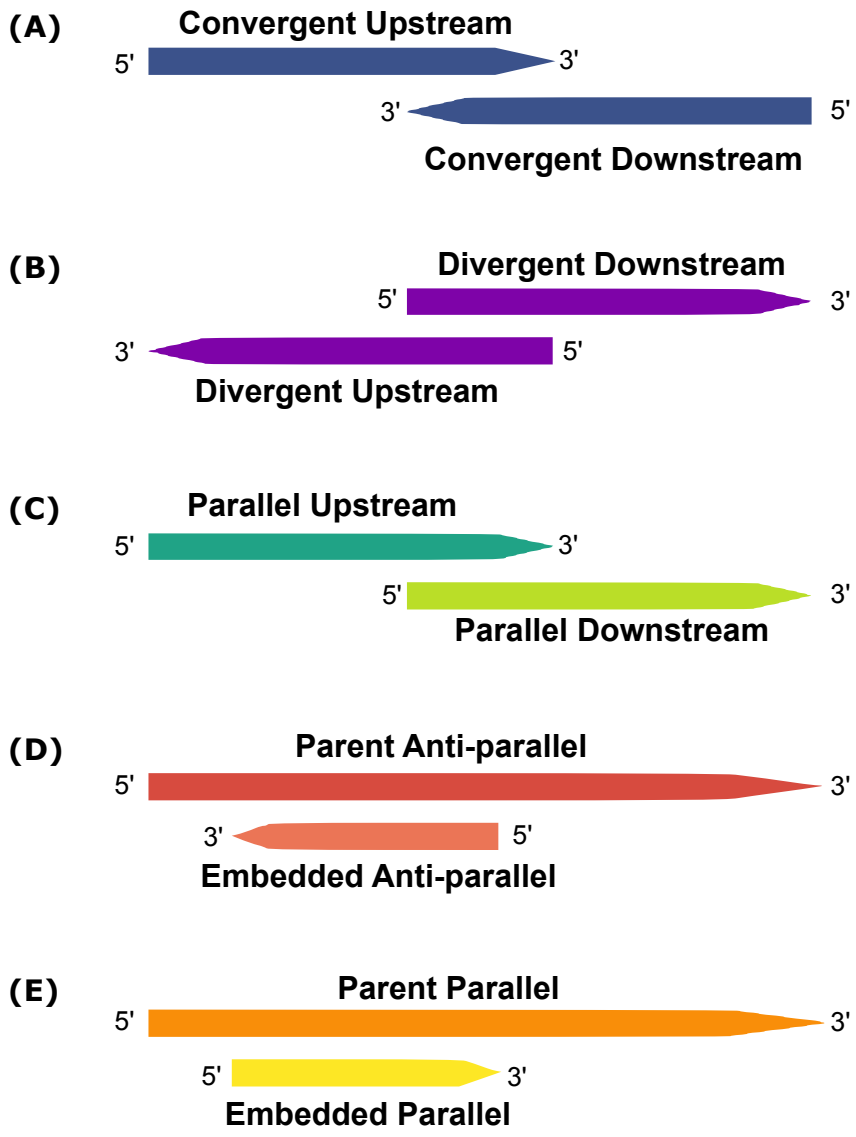
Figure **3-3-2**: Evolutionary event spectrum of divergent and parallel overlapping pairs. The X-axis indicates overlapping pairs, and the Y-axis indicates branches on the 12-fly phylogeny tree. Different colors represent different evolutionary events and black indicates that the overlapping pair has no change on the branch.

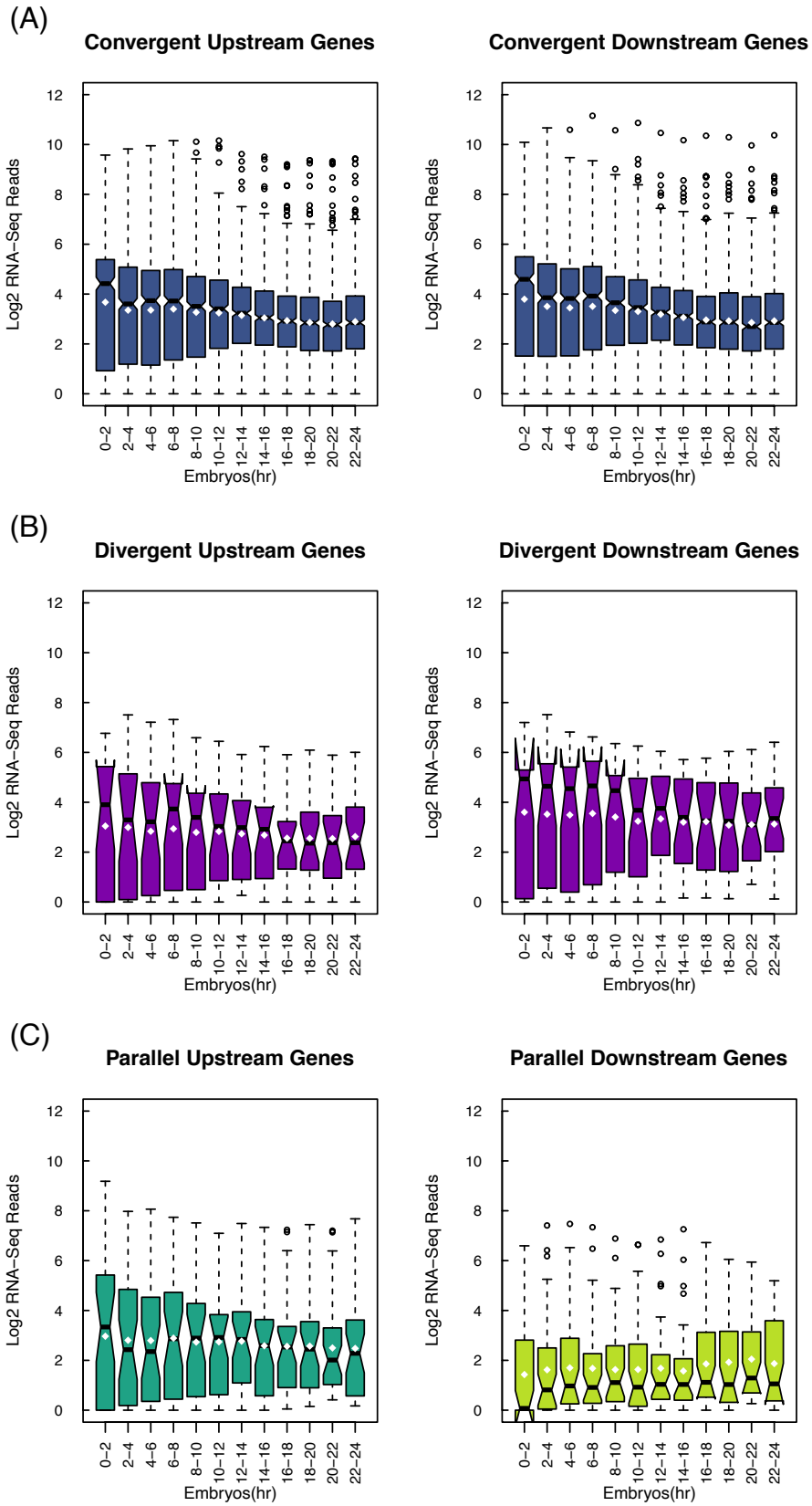
3.4 Overlapping genes display differential expression during *Drosophila* embryonic development.

We tested whether the transcriptional context leads to differential expression for upstream and downstream genes of overlapping pairs. We compared their expression both between and within overlapping pairs for each overlapping type from three perspectives: overall

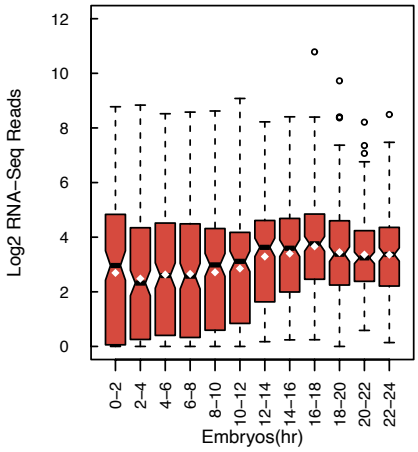
median expression levels, temporal change of median expression pattern and expression variation (Figure **3-4-1**). First, overlapping genes and parental genes in general have higher expression values throughout development regardless of their position. In contrast, both parallel and anti-parallel embedded genes are expressed at a level lower than the parental gene. Second, while convergent and divergent overlapping pairs have comparable expression patterns, parallel overlapping upstream genes tend to decrease in expression during development and parallel overlapping downstream genes tend to increase during development. Parent genes maintain a stable expression level during development, while the embedded genes' expression generally increases over time. Third, the variance in gene expression is generally large early in development, but it decreases over time for partial overlapping genes and parent genes as the development progress. These patterns indicate that overlapping structure correlates with the mode of gene regulation during fly development.

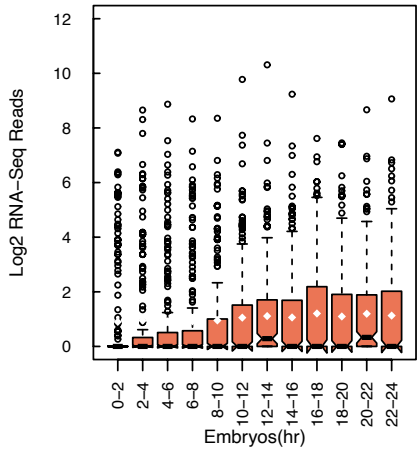
LEGEND


(A)
Convergent Upstream
5'
3'
3'
5'
Convergent Downstream
(B)
Divergent Downstream
5'
3'
3'
5'
Divergent Upstream
(C)
Parallel Upstream
5'
3'
5'
3'
Parallel Downstream
(D)
Parent Anti-parallel
5'
3'
3'
5'
Embedded Anti-parallel
(E)
Parent Parallel
5'
3'
5'
3'
Embedded Parallel

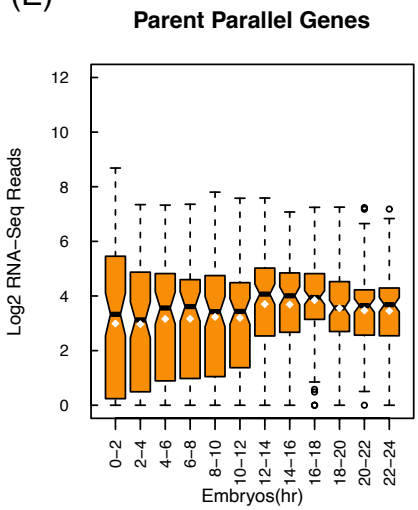

(A)
Convergent Upstream Genes
Convergent Downstream Genes
(B)
Divergent Upstream Genes
Divergent Downstream Genes
(C)
Parallel Upstream Genes
Parallel Downstream Genes
Log2 RNA−Seq Reads
Embryos(hr)
0–2
2–4
4–6
6–8
8–10
10–12
12–14
14–16
16–18
18–20
20–22
22–24

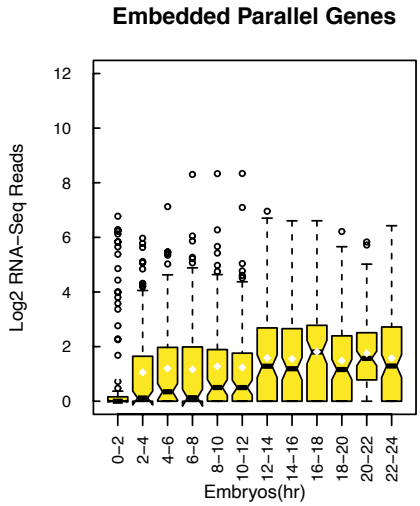
(D)

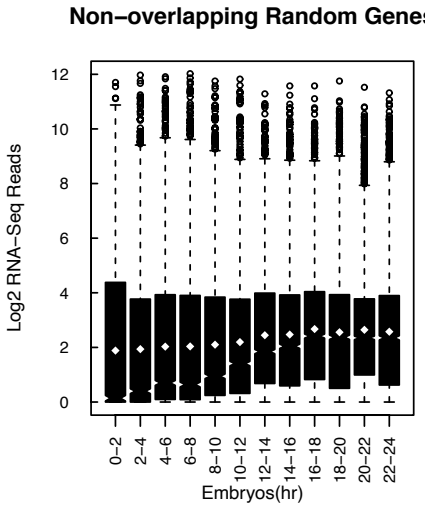

Parent Anti−parallel Genes
Log2 RNA−Seq Reads
0
2
4
6
8
10
12
0−2
2−4
4−6
6−8
8−10
10−12
12−14
14−16
16−18
18−20
20−22
22−24
Embryos(hr)


Embedded Anti−parallel Genes
Log2 RNA−Seq Reads
0
2
4
6
8
10
12
0−2
2−4
4−6
6−8
8−10
10−12
12−14
14−16
16−18
18−20
20−22
22−24
Embryos(hr)

(E)


Parent Parallel Genes
Log2 RNA−Seq Reads
0
2
4
6
8
10
12
0−2
2−4
4−6
6−8
8−10
10−12
12−14
14−16
16−18
18−20
20−22
22−24
Embryos(hr)


Embedded Parallel Genes
Log2 RNA−Seq Reads
0
2
4
6
8
10
12
0−2
2−4
4−6
6−8
8−10
10−12
12−14
14−16
16−18
18−20
20−22
22−24
Embryos(hr)


Non−overlapping Random Genes
Log2 RNA−Seq Reads
0
2
4
6
8
10
12
0−2
2−4
4−6
6−8
8−10
10−12
12−14
14−16
16−18
18−20
20−22
22−24
Embryos(hr)

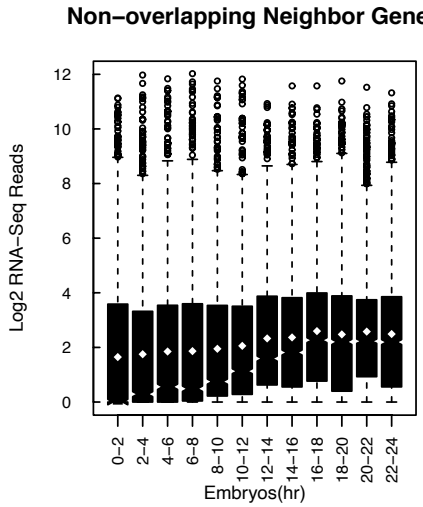

Non−overlapping Neighbor Genes
Log2 RNA−Seq Reads
0
2
4
6
8
10
12
0−2
2−4
4−6
6−8
8−10
10−12
12−14
14−16
16−18
18−20
20−22
22−24
Embryos(hr)

Figure **3-4-1**: Boxplots of overall expression of different overlapping genes. The legend panel denotes the schematic of the groups of overlapping genes in the boxplots. The X-axis indicates 12 developmental stages ordered by time. The Y-axis indicates expression values. The distribution of expression values is summarized with notched boxplot with each color representing one group of overlapping genes: (A) Left: convergent upstream genes expression, Right: convergent downstream genes expression; (B) Left: divergent upstream genes expression, Right: divergent downstream genes expression; (C) Left: parallel upstream genes expression, Right: parallel downstream genes expression; (D) Left: parent anti-parallel genes expression, Right: embedded anti-parallel genes expression; (E) Left: parent parallel genes expression, Right: embedded parallel genes expression; Bottom Left: non-overlapping random genes expression, Bottom Right: non-overlapping neighbor genes expression.

3.5 The interaction between overlapping pairs.

We asked whether overlapping pairs have transcriptional interference between the partners based on the configuration of the overlaps. We estimated the correlation in expression patterns between the overlapping pairs by comparing the Spearman correlation coefficient distribution for the expression values for pairs of genes across the 12 developmental stages (Figure **3-5-1**). We found strong positive correlations between non-overlapping neighboring genes as control comparisons, which is consistent with previous findings where gene clusters are transcriptionally co-regulated (Chen & Stein, 2006; Makalowska et al., 2005; Tsai et al., 2009) We also found unique correlation patterns of overlapping genes: in all five overlapping groups correlations shift towards negative values and the two distributions are significantly different from the non-overlapping neighboring control (Table **3-5-1**). The same strand effect is less than opposite strands, which suggests transcriptional interference between overlapping pairs is more important than the close physical distance between genes.

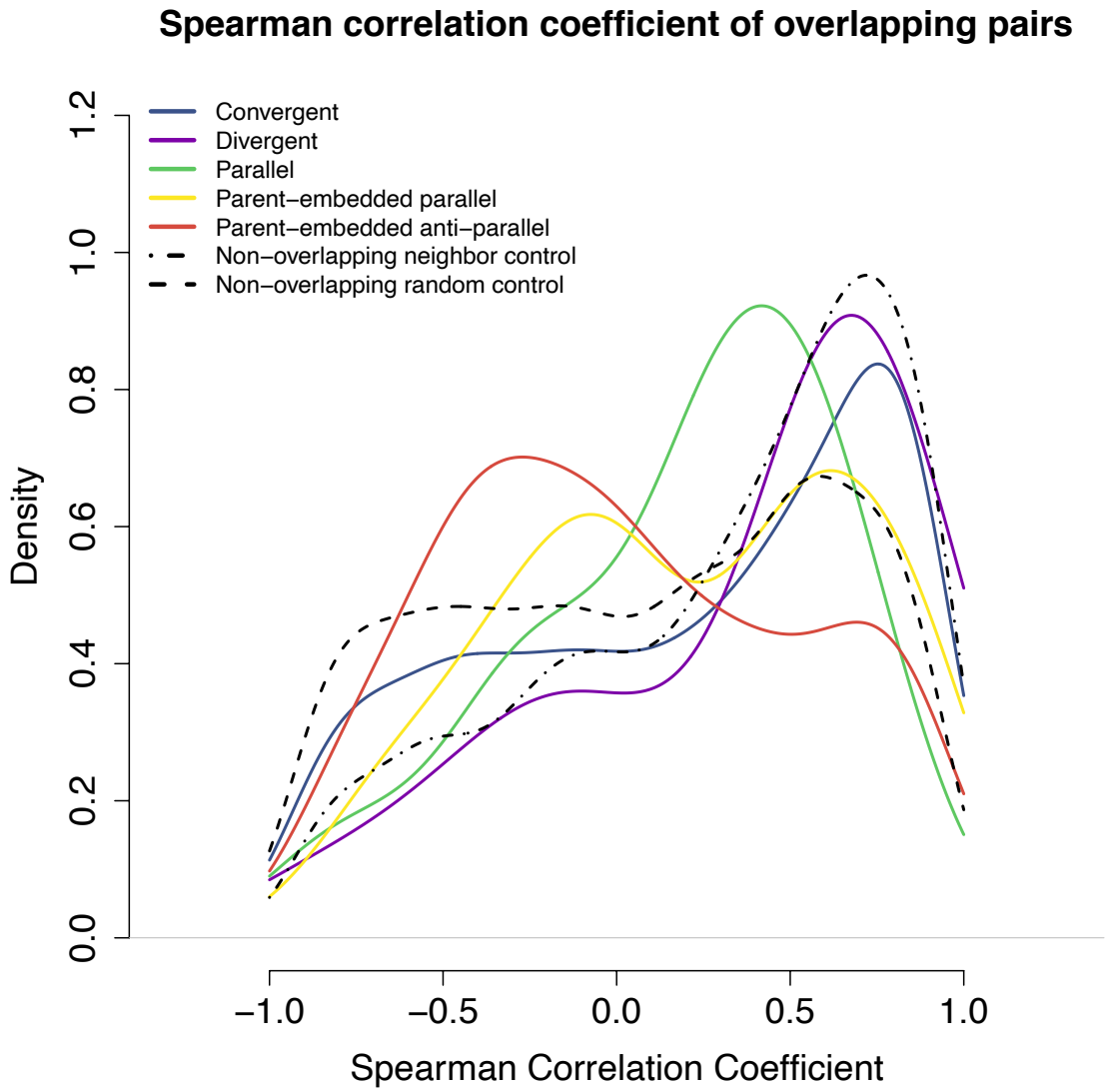


Figure **3-5-1**: Spearman correlation density plot of overlapping pairs. The X-axis indicates spearman correlation coefficient, and the Y-axis shows the density. Non-overlapping random and non-overlapping neighbors are used as the controls.

Table **3-5-1**: Pairwise Kolmogorov–Smirnov Test p-values (p<0.05 with * in the lower left half).

	Convergent	**Divergent**	**Parallel**	**Parent-embedded Parallel**	**Parent-embedded Anti-parallel**	**Non-overlapping Neighbor**	**Non-overlapping Random**
Convergent		1.29E-01	4.60E-02	5.08E-01	1.10E-04	7.64E-04	8.20E-07
Divergent			2.51E-02	8.34E-02	1.31E-03	3.28E-01	8.40E-03
Parallel	*	*		4.43E-01	2.35E-02	1.02E-02	2.60E-01
Parent-embedded Parallel					4.69E-02	3.05E-02	7.41E-02
Parent-embedded Anti-parallel	*	*	*	*		2.40E-09	3.71E-02
Non-overlapping Neighbor	*		*	*	*		2.20E-16
Non-overlapping Random	*	*			*	*	

On the protein level, it has been hypothesized that there might be functional relatedness between overlapping genes. One example in drosophila is a parent-embedded pair, *fig* and *kayak*, belonging to the same cell signal pathway (Hudson et al., 2007). We examined the ontology of the most positively correlated (2.5%) gene pairs and most negatively correlated gene pairs and found no evidence of functional relationship (Table **S2**, Table **S3**).

Chapter 4

Discussion

To answer the overarching question whether overlapping genes are neutral or selected in eukaryotic genomes, we tested hypotheses on the evolutionary level and on the expression level respectively, i.e., whether overlapping gene structure are more conserved than non-overlapping genes in related species and whether overlapping genes express differently under positive effect of transcriptional co-regulation and negative effect of transcription interference. Our analysis (Table **4-1-1**) showed that overlapping genes have distinctive evolutionary and expression patterns compared to non-overlapping genes. Moreover, different overlapping types also behave differently, and these variations correlate with the overlapping proportion and orientation. Noticeably, partial UTR overlaps are in general more abundant, more conserved, and transcriptional coupled while fully embedded overlaps are less frequent, less conserved and transcriptionally decoupled. Opposite strand overlaps significantly outnumber the same strand overlaps and are in general more conserved. The strand effect on the transcription correlation is different for partial and full overlapping structures. In the partial case, same strand overlaps are less correlated in their expression while in the full embedded case, opposite strand overlaps show more negative correlations in expression. Together, we showed that overlapping genes are not a neutral phenomenon and our results support the non-random distribution hypothesis.

Table **4-1-1**: Summary of results from Chapter 3.

Types		**Convergent**		**Divergent**		**Parallel**		**Parent-embedded Anti-parallel**		**Parent-embedded Parallel**		**Non-Overlapping Neighbor Control**	
Overlapping Proportion		Partial						Full				NA	
Overlapping Orientation		Opposite				Same		Opposite		Same			
Overlapping Seq Context		Antisense 3'-3'		Antisense 5'-5'		Sense 3'-5'		Antisense		Sense			
Number	**Expected Ratio**	1 : 1 : 2						1 : 1					
	Observed Number	787		25		40		207		112			
Conservation Degree		72.32 ±3.19%		28.00 ±17.60%		57.50 ±15.32%		48.79 ±6.81%		23.19 ±9.96%		34.05 ±2.24%	
Expression Level		H	H	H	H	H	L	H	L	H	L	M	M
Trend Over Developmental Course		↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗	↗
Expression Correlation Shifting Compared to Non-Overlapping Neighbor Control		-		-		- -		- - -		- -		+	

Note: H: High, L: Low, "-": negative shift, "+": positive expression correlation.

In this study, we found over 30% or protein coding genes overlap in *D. melanogaster*. The number of overlapping genes in the *D. melanogaster* genome are higher than previous findings based on all the other eukaryotic organism (Table **1-3-1**). Genome size and gene density cannot fully explain it. The genome size of *D. melanogaster* is around 180Mb, and *D. melanogaster* has an average gene density of 1/13,000 bp compared to around 1/148,000bp in humans. However, *C. elegans* has a more compact genome than *D. melanogaster* and still only

have 12% of genes that overlap. Alternatively, the data source and quality variation in earlier research might be the reason. Increasing new protein gene models are annotated as genomic data improve over the years, as showed in human: a revisit of human overlapping protein-coding genes (Chen et al., 2019) found a significantly higher overlapping percentage 25% compared to 14% in previous studies (Veeramachaneni et al., 2004). Furthermore, we found the distribution of overlapping types deviates significantly from the random expectation. Under a neutral model, the number of convergent overlaps, divergent overlaps and parallel overlaps should follow 1:1:2 ratio and the number of parent-embedded sense and antisense overlaps should follow 1:1 ratio. However, we found an overrepresentation of 3'-3' overlaps and parent-embedded antisense overlaps, and an underrepresentation of 5'-5', 3-5' and parent-embedded sense overlaps (Figure **3-1-2**, Table **4-1-1**). For the partial overlap case, it might be due to strong selection against overlapping involving 5' end of gene which contains the promoter region. For the parent-embedded case, Lee & Chang (2013) proposed that the underrepresentation of same strand overlaps is likely caused by selection against mis-splicing because they found the parallel embedded genes are more likely to be single exon genes and in the multi-exon case still have fewer exons than anti-parallel counter partners. These deviations show the distribution of overlapping genes on the genome are not random and suggest overlapping genes may subject to different selection pressures.

Our evolutionary analysis using comparative methods of overlapping genes across the 12 fly species gives another piece of evidence that different overlapping structures are subject to different selection pressures (Table **4-1-1**). Partial overlaps tend to be more conserved than full overlaps whereas opposite strand overlaps tend to be more conserved than the same strand overlaps (the only exception is the divergent overlaps). In the extreme case of parent-embedded parallel case the conservation degree is marginally lower than the non-overlapping controls. Comparing the outlier divergent overlap with the parent-embedded parallel overlap, we found

although both types of overlaps show poor conservation across the species, and they have quite different evolutionary history. A detailed look into divergent structure history reveals that gene rearrangements are the leading mechanism responsible for divergent structure formation, which occur mostly in the branch leading to the speciation of melanogaster subgroups. On the other hand, *D. melanogaster* specific gene gain explains most of the low conservation of parent-embedded parallel overlaps in all the other non-*D. melanogaster* species. In fact, gene birth within the intron of the other gene might be a way the genomes acquire the complexity along evolution (Assis et al., 2008). At the time that this research was originally performed, the non-melanogaster species did not have UTR annotation and thus it was impossible to infer transcript overlap, we used *D. melanogaster* as our reference species and all the overlapping genes were taken from *D. melanogaster* genome. A pairwise all versus all comparisons in each individual species in the future would give a more accurate estimation of conservation degrees, which would help differentiate the effect of reference species specific evolution from gradual evolution in all the other non-reference species. Annotation of UTRs has improved since 2015, which would also help to rigorously test for conservation of overlapping gene structure.

Our expression analysis on different overlapping groups suggested that gene expression might have played a role affecting the conservation of overlapping genes. We compared the expression levels of overlapping genes in the *D. melanogaster* embryos from different developmental stages (Table **4-1-1**). First, both partner genes of overlapping pairs, expect for embedded genes, tend to have high expression values than non-overlapping genes. There are multiple pieces of evidence for coordinated expression for neighboring gene clusters in general (Eisen et al., 1998; Li et al., 2010; Tsai et al., 2009). And this clustering effect could be caused by coregulation of genes residing within topologically associating domains (TADs). (Fraser et al., 2015). Second, in contrast to non-overlapping genes and parent genes, convergent overlapping genes, divergent overlapping genes, parallel upstream genes tend to be highly expressed in the

early development compared to non-overlapping genes suggest they may play vital roles in early fly developmental, which might explain why these genes are more conserved than non-overlapping genes. On the other hand, embedded genes and parallel downstream genes are only highly expressed at later stages of development suggesting less selection constraint during development. Moreover, it has been suggested that there is a parallel relationship between ontogeny and phylogeny at expression level. Older genes are expressed in early phases while newly evolved genes are expressed at later stages of development. (Domazet-Lošo & Tautz, 2010). According to this model, this is indeed consistent with the conservation level of genes. Most convergent overlapping genes are conserved in all fly species and arise before the common ancestor of the twelve species and are retained in all species ever since. Embedded genes are more recently gained than their parent genes, this finding has been consistent with previous research about embedded genes. (Henikoff et al., 1986; Lee & Chang, 2013; Yu et al., 2005)

We asked whether overlapping pairs have transcriptional interference or functionally relatedness between the partners based on the configuration of the overlaps and our results support the former while no evidence was found for the latter between the overlapping pairs on the genome level. We showed that the emergence of the second transcript shifts the expression correlation of the overlapping pair towards less correlated side in all groups and this effect is more significant in the parent-embedded pairs (Table **4-1-1**), which suggests overlapping transcripts may override the effect of chromatin coregulation and induce negative correlation in overlapping gene expression. The shifting in partial overlaps is slight, suggesting transcription interference caused by potential entanglement of UTR region, either through antisense regulation (Fire et al., 1998) or transcription collision (Assis, 2016; Crampton et al., 2006), is minimum. However, 5’ end overlaps seem to be more severely affected than 3’ end overlaps, suggesting promoter region near the 5’ end region of the downstream gene causes more transcriptional decoupling. Compared to partial overlaps, full overlaps between parents and embedded genes

which have complementary sequences and completely overlap tend to be more transcriptionally decoupled. This negative correlated pattern might be an effect of intronic genes interfering with transcription of its parent gene by disturbing normal proper splicing process (Kaer et al., 2011) and inducing mis-splicing (Lee & Chang, 2013). The antisense overlap is the most negatively correlated of all, suggesting intronic genes might have regulatory functions over the parent genes through RNAi or a coping mechanism avoiding spacetime co-expression (Henikoff et al., 1986). Overall, our findings on full overlapping parent-embedded pairs are generally consistent with studies in parent-embedded genes in *C. elegans* (Chen & Stein, 2006), *D. melanogaster* (Lee & Chang, 2013) and humans (Yu et al., 2005) and in mammals (Assis, 2016). With new transcriptome data available for the genus *Drosophila* (Li et al., 2022), we could be doing expression correlation analysis in the future for all the non-*D. melanogaster* species to see if the similar pattern holds (Table **4-1-2**).

Table **4-1-2**: New RNA-seq data of the 11 non-*D. melanogaster* species. (Source: NCBI)

NCBI Project ID	Samples	Sequencing Tech	Submission Date
PRJNA314145	Dmoj	RNA-seq	03_02_16
PRJNA317989	Dgri: five strains	RNA-seq	04_11_16
PRJNA326536	Dpse	RNA-seq	06_22_16
PRJNA337934	Dana Dere, Dsec, Dsim, Dvir	RNA-seq	08_05_16
PRJNA351832	Dsim, Dere, Dana, Dpse: embryos	RNA-seq	10_28_16
PRJNA376405	Dvir: male reproductive tissue and gonadectomized carcass	RNA-seq	02_22_17
PRJNA388952	Dyak, Dana, Dpse, Dper, Dwil, Dmoj, Dvir	RNA-seq	06_01_17
PRJNA393203	Dyak, Dpse, Dsim, Dper	RNA-seq	07_05_17
PRJNA395148	Dmoj	RNA-seq	07_20_17
PRJNA420518	Dpse	RNA-seq	11_30_17
PRJNA424164	Dana	RNA-seq	12_20_17
PRJNA449374	the 11 non-Dmel species: embryos from 3–8 day old females	RNA-seq	04_09_18
PRJNA507780	Dsec larvae	RNA-seq	11_30_18
PRJNA548113	Dpse: male testes	RNA-seq	06_10_19
PRJNA553234	Dsim Dyak: female and male brains	RNA-seq	07_08_19
PRJDB6923	Dsim: male genitalia	RNA-seq	08_17_19

PRJNA574480	Dmoj: from the Sonoran Desert and Santa Catalina Island	RNA-seq	09_26_19
PRJNA574647	Dsec, D.sim	RNA-seq	09_27_19
PRJNA575046	Dyak, Dsim	RNA-seq	09_30_19
PRJNA750150	Dana: multinucleated giant hemocytes	RNA-seq	07_27_21
PRJNA764902	Dwil: gonads	RNA-seq	09_21_21
PRJNA767871	Dsim, Dsec	RNA-seq	10_01_21
PRJNA768815	Dsim, Dsec: testes and accessory glands	RNA-seq	10_05_21
PRJNA845784	Dsec: adult female	RNA-seq	06_03_22
PRJNA855483	Dyak, Dsim	RNA-seq	07_04_22
PRJDB13401	Dsec: guts	RNA-seq	10_19_22
PRJNA1041314	Dsim and Dsec: central brains	snRNA-Seq	11_16_23
PRJNA1093075	Dyak: doublesex expressing neurons of males	scRNAseq	03_28_24
PRJDB6950	Dsec: male genitalia	RNA-seq	04_01_24
PRJNA1124160	Dvir: various developmental stages	RNA-seq	06_14_24

Our analysis of overlapping gene conservation in *D. simulans* and *D. sechellia* should be viewed with caution. The deviation of general evolutionary pattern of *D. simulans* and *D. sechellia* could come from the annotation quality difference between *D. melanogaster* and non-*D. melanogaster* species at the time of this study (2013). Overlapping genes are notoriously difficult to annotate due to their sequence sharing nature (Wright et al., 2022). Sequencing qualities of non-*D. melanogaster* species at the time of this research may also have played a role. *D. sechellia* was sequenced to 3x coverage and the initial sequence of *D. simulans* used different strains sequenced from 1x to 5x coverage to create a mosaic assembly. The mosaic assembly of the *D. simulans* data from the different strains was more fragmented due to polymorphism that existed among the strains, which resulted in genes incorrectly dropped from scaffolds (Schaeffer et al., 2008). *D. simulans* assemblies got worse as data from additional strains was added. It was thought at the time that one could get a genome and nucleotide polymorphism data simultaneously, however, this was not the case (Begun et al., 2007). This sequencing issue cannot be remedied by an annotation quality check. Although we did an annotation artifacts analysis on

D. sechellia and *D. simulans* to remove possible artifacts in overlapping ortholog identification, these two species still show lower than expected degree of conservation and higher than expected evolutionary events leading to overlap structure loss on their lineages, respectively. With recent available high coverage Drosophila sequences and new annotations (Kim et al., 2021, Tvedte et al., 2021, Li et al., 2022, Kim et al., 2024, Table **4-1-3**). This caveat should be resolved.

Table **4-1-3**: Sequence, assembly and annotation updates in the 12 fly species up until now. (Source: NCBI)

Species	Submission Date	Annotation Date	Assembly Level	Coverage	Seq Tech	NCBI RefSeq Assembly
Dana	07.16.21	10.14.21	Chromosome	370.0x	PacBio Sequel II; Oxford Nanopore MinION	GCF_017639315.1
Dere	10.23.18	11.01.18	Contig	190.0x	PacBio Sequel; Illumina	GCF_003286155.1
Dgri	04.28.21	08.19.21	Contig	92.8x	Oxford Nanopore MinION; Illumina HiSeq	GCF_018153295.1
Dmoj	04.28.21	09.24.21	Contig	120.4x	Oxford Nanopore MinION; Illumina HiSeq	GCF_018153725.1
Dper	10.23.18	11.05.18	Contig	130.0x	PacBio Sequel; Illumina	GCF_003286085.1
Dpse	03.03.20	04.08.20	Chromosome	280.0x	PacBio Sequel	GCF_009870125.1
Dsec	04.16.19	03.03.20	Chromosome	115.0x	PacBio	GCF_004382195.2
Dsim	10.05.21	10.28.21	Chromosome	120.0x	PacBio RSII; Illumina HiSeq	GCF_016746395.2
Dvir	10.23.18	02.04.20	Contig	100.0x	PacBio Sequel; Illumina	GCF_003285735.1
Dwil	06.16.21	03.01.22	Chromosome	150.0x	PacBio RS; Illumina	GCF_018902025.1
Dyak	07.30.21	08.16.21	Chromosome	80.0x	PacBio RSII; Illumina HiSeq	GCF_016746365.2

In conclusion, overlapping genes are a common phenomenon in fruit flies. They are generally more conserved in their arrangement than non-overlapping genes and the conservation degree varies with overlapping configuration, namely relative orientation and overlapping

proportion. Different overlapping configuration also show different evolutionary pattern with gene gain/loss being the dominating mechanism. We showed that overlapping genes are not a neutral phenomenon and our results support the non-random distribution hypothesis. We showed other than physical cluster effect, overlapping could be another factor, which have distinctive expression and evolutionary pattern than gene clusters. Finally, our results support the new standpoint of view that DNA sequence could have multiple functions carrying both coding and regulation function.

Data Availability: Supplementary data and computer codes are available at https://scholarsphere.psu.edu/resources/d3c340a6-4316-4bbe-9ba2-e483aaf2c3af.

Bibliography

Assis, R. (2016). Transcriptional interference promotes rapid expression divergence of drosophila nested genes. *Genome Biology and Evolution*, *8*(10), 3149–3158. https://doi.org/10.1093/gbe/evw237

Assis, R., Kondrashov, A. S., Koonin, E. V, & Kondrashov, F. A. (2008). Nested genes and increasing organizational complexity of metazoan genomes. *Trends in Genetics*, *24*(10), 475–478. https://doi.org/10.1016/j.tig.2008.08.003

Barrell, B. G., Air, G. M., & Hutchison, C. a. (1976). Overlapping genes in bacteriophage phiX174. *Nature*, *264*(5581), 34–41. https://doi.org/10.1038/264034a0

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. D. W., Poh, Y. P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., & Langley, C. H. (2007). Population genomics: Whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biology*, *5*(11), 2534–2559. https://doi.org/10.1371/journal.pbio.0050310

Behura, S. K., & Severson, D. W. (2013). Overlapping genes of Aedes aegypti: evolutionary implications from comparison with orthologs of Anopheles gambiae and other insects. *BMC Evolutionary Biology*, *13*(1), 124. https://doi.org/10.1186/1471-2148-13-124

Belshaw, R., Pybus, O. G., & Rambaut, A. (2007). The evolution of genome compression and genomic novelty in RNA viruses. *Genome Research*, *17*(10), 1496–1504. https://doi.org/10.1101/gr.6305707

Bhutkar, A., Schaeffer, S. W., Russo, S. M., Xu, M., Smith, T. F., & Gelbart, W. M. (2008). Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes. *Genetics*, *179*(3), 1657–1680. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18622036

Bukhnikashvili, L. (2023). Overlaps Between CDS Regions of Protein-Coding Genes in the Human Genome: A Case Study on the NR1D1-THRA Gene Pair. *Journal of Molecular Evolution*, *91*(6), 963–975. https://doi.org/10.1007/s00239-023-10147-8

Calvete, O., González, J., Betrán, E., & Ruiz, A. (2012). Segmental duplication, microinversion, and gene loss associated with a complex inversion breakpoint region in drosophila. *Molecular Biology and Evolution*, *29*(7), 1875–1889. https://doi.org/10.1093/molbev/mss067

Chang, C. H., Chavan, A., Palladino, J., Wei, X., Martins, N. M. C., Santinello, B., Chen, C. C., Erceg, J., Beliveau, B. J., Wu, C. T., Larracuente, A. M., & Mellone, B. G. (2019). Islands of retroelements are major components of Drosophila centromeres. In *PLoS Biology* (Vol. 17, Issue 5). https://doi.org/10.1371/journal.pbio.3000241

Chen, C. H., Pan, C. Y., & Lin, W. chang. (2019). Overlapping protein-coding genes in human genome and their coincidental expression in tissues. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-49802-w

Chen, J., Brunner, A.-D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., Itzhak, D. N., Li, J. Y., Mann, M., Leonetti, M. D., & Weissman, J. S. (2020). Pervasive functional translation of noncanonical human open reading frames. *Science*, *367*(6482), 1140 LP – 1146. https://doi.org/10.1126/science.aay0262

Chen, N., & Stein, L. D. (2006). Conservation and functional significance of gene topology in the genome of Caenorhabditis elegans. *Genome Research*, *16*(5), 606–617. https://doi.org/10.1101/gr.4515306

Chirico, N., Vianelli, A., & Belshaw, R. (2010). Why genes overlap in viruses. *Proceedings. Biological Sciences / The Royal Society*, *277*(1701), 3809–3817. https://doi.org/10.1098/rspb.2010.1052

Chlebek, J. L., Leonard, S. P., Kang-Yun, C., Yung, M. C., Ricci, D. P., Jiao, Y., & Park, D. M. (2023). Prolonging genetic circuit stability through adaptive evolution of overlapping genes. *Nucleic Acids Research*, *51*(13), 7094–7108. https://doi.org/10.1093/nar/gkad484

Consortium, D. 12 G., Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., Pollard, D. A., Sackton, T. B., Larracuente, A. M., Singh, N. D., Abad, J. P., Abt, D. N., Adryan, B., Aguade, M., … MacCallum, I. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, *450*(7167), 203–218. https://doi.org/10.1038/nature06341

Crampton, N., Bonass, W. A., Kirkham, J., Rivetti, C., & Thomson, N. H. (2006). Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. *Nucleic Acids Research*, *34*(19), 5416–5425. https://doi.org/10.1093/nar/gkl668

Cremer, T., & Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, *2*(4), 292–301. https://doi.org/10.1038/35066075

Dan, I., Watanabe, N. M., Kajikawa, E., Ishida, T., Pandey, A., & Kusumi, A. (2002). Overlapping of MINK and CHRNE gene loci in the course of mammalian evolution. *Nucleic Acids Research*, *30*(13), 2906–2910. https://doi.org/10.1093/nar/gkf407

Domazet-Lošo, T., & Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, *468*(7325), 815–818. https://doi.org/10.1038/nature09632

Eagen, K. P., Aiden, E. L., & Kornberg, R. D. (2017). Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(33), 8764–8769. https://doi.org/10.1073/pnas.1701291114

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(25), 14863–14868. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=24541&tool=pmcentrez&rendertype=abstract

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, *391*(6669), 806–811. https://doi.org/10.1038/35888

Fraser, J., Williamson, I., Bickmore, W. A., & Dostie, J. (2015). An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiology and Molecular Biology Reviews*, *79*(3), 347–372. https://doi.org/10.1128/mmbr.00006-15

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., Brown, J. B., Cherbas, L., Davis, C. A., Dobin, A., Li, R., Lin, W., Malone, J. H., Mattiuzzo, N. R., Miller, D., … Celniker, S. E. (2011). The developmental transcriptome of Drosophila melanogaster. *Nature*, *471*(7339), 473–479. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=21179090

Henikoff, S., Keene, M. A., Fechtel, K., & Fristrom, J. W. (1986). Gene within a gene: nested Drosophila genes encode unrelated proteins on opposite DNA strands. *Cell*, *44*(1), 33–42. http://www.ncbi.nlm.nih.gov/pubmed/3079672

Hou, C., Li, L., Qin, Z. S., & Corces, V. G. (2012). Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Molecular Cell*, *48*(3), 471–484. https://doi.org/10.1016/j.molcel.2012.08.031

Hudson, S. G., Garrett, M. J., Carlson, J. W., Micklem, G., Celniker, S. E., Goldstein, E. S., & Newfeld, S. J. (2007). Phylogenetic and genomewide analyses suggest a functional relationship between kayak, the Drosophila fos homolog, and fig, a predicted protein phosphatase 2c nested within a kayak intron. *Genetics*, *177*(3), 1349–1361. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18039871

Hug, C. B., Grimaldi, A. G., Kruse, K., & Vaquerizas, J. M. (2017). Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell*, *169*(2), 216-228.e19. https://doi.org/10.1016/j.cell.2017.03.024

Inagaki, S., Takahashi, M., Takashima, K., Oya, S., & Kakutani, T. (2021). Chromatin-based mechanisms to coordinate convergent overlapping transcription. *Nature Plants*, *7*(3), 295–302. https://doi.org/10.1038/s41477-021-00868-3

Johnson, Z. I., & Chisholm, S. W. (2004). Properties of overlapping genes are conserved across microbial genomes. *Genome Research*, *14*(11), 2268–2272. https://doi.org/10.1101/gr.2433104

Kaer, K., Branovets, J., Hallikma, A., Nigumann, P., & Speek, M. (2011). Intronic L1 Retrotransposons and Nested Genes Cause Transcriptional Interference by Inducing Intron Retention, Exonization and Cryptic Polyadenylation. *PLoS ONE*, *6*(10), e26099. https://doi.org/10.1371/journal.pone.0026099

Keese, P. K., & Gibbs, A. (1992). Origins of genes: “Big bang” or continuous creation? *Proceedings of the National Academy of Sciences of the United States of America*, *89*(20), 9489–9493. https://doi.org/10.1073/pnas.89.20.9489

Kharchenko, P. V, Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M. Y., Luquette, L. J., Xi, R., Jung, Y. L., Park, R. W., Bishop,

E. P., … Park, P. J. (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, *471*(7339), 480–485. https://doi.org/10.1038/nature09725

Kim, B. Y., Gellert, H. R., Church, S. H., Suvorov, A., Anderson, S. S., Barmina, O., Beskid, S. G., Comeault, A. A., Crown, K. N., Diamond, S. E., Dorus, S., Fujichika, T., Hemker, J. A., Hrcek, J., Kankare, M., Katoh, T., Magnacca, K. N., Martin, R. A., Matsunaga, T., … Petrov, D. A. (2024). Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. *PLoS Biology*, *22*(7 JULY), 1–23. https://doi.org/10.1371/journal.pbio.3002697

Kim, B. Y., Wang, J. R., Miller, D. E., Barmina, O., Delaney, E., Thompson, A., Comeault, A. A., Peede, D., D'agostino, E. R. R., Pelaez, J., Aguilar, J. M., Haji, D., Matsunaga, T., Armstrong, E. E., Zych, M., Ogawa, Y., Stamenković-Radak, M., Jelić, M., Veselinović, M. S., … Petrov, D. A. (2021). Highly contiguous assemblies of 101 drosophilid genomes. *ELife*, *10*, 1–32. https://doi.org/10.7554/eLife.66405

Kumar, S., Stecher, G., Suleski, M., & Blair Hedges, S. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/MOLBEV/MSX116

Lee, Y. C. G., & Chang, H. H. (2013). The evolution and functional significance of nested gene structures in Drosophila melanogaster. *Genome Biology and Evolution*, *5*(10), 1978–1985. https://doi.org/10.1093/gbe/evt149

Lercher, M. J., Urrutia, A. O., & Hurst, L. D. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genetics*, *31*(2), 180–183. https://doi.org/10.1038/ng887

Li, F., Rane, R. V., Luria, V., Xiong, Z., Chen, J., Li, Z., Catullo, R. A., Griffin, P. C., Schiffer, M., Pearce, S., Lee, S. F., McElroy, K., Stocker, A., Shirriffs, J., Cockerell, F., Coppin, C., Sgrò, C. M., Karger, A., Cain, J. W., … Zhang, G. (2022). Phylogenomic analyses of the

genus Drosophila reveals genomic signals of climate adaptation. *Molecular Ecology Resources*, *22*(4), 1559–1581. https://doi.org/10.1111/1755-0998.13561

Li, S., Shih, C.-H., & Kohn, M. H. (2010). Functional and evolutionary correlates of gene constellations in the Drosophila melanogaster genome that deviate from the stereotypical gene architecture. *BMC Genomics*, *11*(1), 322. https://doi.org/10.1186/1471-2164-11-322

Liao, Y., Zhang, X., Chakraborty, M., & Emerson, J. J. (2021). Topologically associating domains and their role in the evolution of genome structure and function in Drosophila. *Genome Research*, *31*(3), 397–410. https://doi.org/10.1101/GR.266130.120

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

Makałowska, I., Lin, C.-F., & Hernandez, K. (2007). Birth and death of gene overlaps in vertebrates. *BMC Evolutionary Biology*, *7*(1), 193. https://doi.org/10.1186/1471-2148-7-193

Makalowska, I., Lin, C.-F., & Makalowski, W. (2005). Overlapping genes in vertebrate genomes. *Computational Biology and Chemistry*, *29*(1), 1–12. https://doi.org/10.1016/j.compbiolchem.2004.12.006

Munroe, S. H., Morales, C. H., Duyck, T. H., & Waters, P. D. (2015). Evolution of the antisense overlap between genes for thyroid hormone receptor and rev-erbα and characterization of an exonic G-rich element that regulates splicing of TRα2 mRNA. *PLoS ONE*, *10*(9), 1–25. https://doi.org/10.1371/journal.pone.0137893

Passalacqua, K. D., Varadarajan, A., Weist, C., Ondov, B. D., Byrd, B., Read, T. D., & Bergman, N. H. (2012). Strand-Specific RNA-Seq Reveals Ordered Patterns of Sense and Antisense

Transcription in Bacillus anthracis. *PLoS ONE*, *7*(8), e43350. https://doi.org/10.1371/journal.pone.0043350

Pavesi, A. (2006). Origin and evolution of overlapping genes in the family Microviridae. *Journal of General Virology*, *87*(4), 1013–1017. https://doi.org/10.1099/vir.0.81375-0

Pavesi, A., Vianelli, A., Chirico, N., Bao, Y., Blinkova, O., Belshaw, R., Firth, A., & Karlin, D. (2018). Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLoS ONE*, *13*(10), 1–24. https://doi.org/10.1371/journal.pone.0202513

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A., & Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-017-02525-w

Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M. J., Chen, R., Meisel, R. P., Couronne, O., Hua, S., Smith, M. A., Zhang, P., Liu, J., Bussemaker, H. J., van Batenburg, M. F., Howells, S. L., Scherer, S. E., … Gibbs, R. A. (2005). Comparative genome sequencing of Drosophila pseudoobscura : Chromosomal, gene, and cis -element evolution. *Genome Research*, *15*(1), 1–18. https://doi.org/10.1101/gr.3059305

Rogozin, I. B., Spiridonov, A. N., Sorokin, A. V, Wolf, Y. I., Jordan, I. K., Tatusov, R. L., & Koonin, E. V. (2002). Purifying and directional selection in overlapping prokaryotic genes. *Trends in Genetics*, *18*(5), 228–232. https://doi.org/10.1016/S0168-9525(02)02649-5

Sakharkar, K. R., Sakharkar, M. K., Verma, C., & Chow, V. T. K. (2005). Comparative study of overlapping genes in bacteria, with special reference to Rickettsia prowazekii and Rickettsia conorii. *International Journal of Systematic and Evolutionary Microbiology*, *55*(Pt 3), 1205–1209. https://doi.org/10.1099/ijs.0.63446-0

Sankoff, D. (1975). Minimal Mutation Trees of Sequences. *SIAM Journal on Applied Mathematics*, *28*(1), 35–42. https://doi.org/10.1137/0128004

Sanna, C. R., Li, W.-H., & Zhang, L. (2008). Overlapping genes in the human and mouse genomes. *BMC Genomics*, *9*(1), 169. https://doi.org/10.1186/1471-2164-9-169

Schaeffer, S. W. (2018). Muller "Elements" in Drosophila : How the Search for the Genetic Basis for Speciation Led to the Birth of Comparative Genomics. *Genetics*, *210*(1), 3–13. https://doi.org/10.1534/genetics.118.301084

Schaeffer, S. W., Bhutkar, A., McAllister, B. F., Matsuda, M., Matzkin, L. M., O'Grady, P. M., Rohde, C., Valente, V. L., Aguade, M., Anderson, W. W., Edwards, K., Garcia, A. C., Goodman, J., Hartigan, J., Kataoka, E., Lapoint, R. T., Lozovsky, E. R., Machado, C. A., Noor, M. A., … Kaufman, T. C. (2008). Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. *Genetics*, *179*(3), 1601–1655. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18622037

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L., & Ren, B. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports*, *17*(8), 2042–2059. https://doi.org/10.1016/j.celrep.2016.10.061

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., & Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, *148*(3), 458–472. https://doi.org/10.1016/j.cell.2012.01.010

Shintani, S., O'hUigin, C., Toyosawa, S., Michalová, V., & Klein, J. (1999). Origin of gene overlap: the case of TCP1 and ACAT2. *Genetics*, *152*(2), 743–754.

Siegel, T., Hon, C.-C., Zhang, Q., Lopez-Rubio, J.-J., Scheidig-Benatar, C., Martins, R. M., Sismeiro, O., Coppée, J.-Y., & Scherf, A. (2014). Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. *BMC Genomics*, *15*(1), 150. https://doi.org/10.1186/1471-2164-15-150

Soldà, G., Suyama, M., Pelucchi, P., Boi, S., Guffanti, A., Rizzi, E., Bork, P., Tenchini, M. L., & Ciccarelli, F. D. (2008). Non-random retention of protein-coding overlapping genes in Metazoa. *BMC Genomics*, *9*, 1–12. https://doi.org/10.1186/1471-2164-9-174

Spellman, P. T., & Rubin, G. M. (2002). Evidence for large domains of similarly expressed genes in the Drosophila genome. *Journal of Biology*, *1*(1), 5. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12144710

Thygesen, H. H., & Zwinderman, A. H. (2005). Modelling the correlation between the activities of adjacent genes in Drosophila. *BMC Bioinformatics*, *6*, 10. https://doi.org/10.1186/1471-2105-6-10

Tsai, H.-K., Huang, P.-Y., Kao, C.-Y., & Wang, D. (2009). Co-Expression of Neighboring Genes in the Zebrafish (Danio rerio) Genome. *International Journal of Molecular Sciences*, *10*(8), 3658–3670. https://doi.org/10.3390/ijms10083658

Tvedte, E. S., Gasser, M., Sparklin, B. C., Michalski, J., Hjelmen, C. E., Johnston, J. S., Zhao, X., Bromley, R., Tallon, L. J., Sadzewicz, L., Rasko, D. A., & Dunning Hotopp, J. C. (2021). Comparison of long-read sequencing technologies in interrogating bacteria and fly genomes. *G3: Genes, Genomes, Genetics*, *11*(6). https://doi.org/10.1093/g3journal/jkab083

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R., & Zhang, H. (2009). FlyBase: enhancing

Drosophila Gene Ontology annotations. *Nucleic Acids Research*, *37*(Database), D555–D559. https://doi.org/10.1093/nar/gkn788

Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R., & Makalowska, I. (2004). Mammalian Overlapping Genes: The Comparative Perspective. *Genome Research*, *14*(2), 280–286. https://doi.org/10.1101/gr.1590904

Wadhawan, S., Dickins, B., & Nekrutenko, A. (2008). Wheels within Wheels: Clues to the Evolution of the Gnas and Gnal Loci. *Molecular Biology and Evolution*, *25*(12), 2745–2757. https://doi.org/10.1093/molbev/msn229

Wright, B. W., Molloy, M. P., & Jaschke, P. R. (2022). Overlapping genes in natural and engineered genomes. *Nature Reviews Genetics*, *23*(3), 154–168. https://doi.org/10.1038/s41576-021-00417-w

Wright, D., & Schaeffer, S. W. (2022). The relevance of chromatin architecture to genome rearrangements in Drosophila. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1856). https://doi.org/10.1098/rstb.2021.0206

Yu, P., Ma, D., & Xu, M. (2005). Nested genes in the human genome. *Genomics*, *86*(4), 414–422. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16084061