The Evolution and Functional Significance of Nested Gene Structures in *Drosophila melanogaster*

Yuh Chwen G. Lee[1,3,†] and Hsiao-Han Chang[2,4,*,†]

[1]Center for Population Biology and Department of Evolution and Ecology, University of California

[2]Department of Organismic and Evolutionary Biology, Harvard University

[3]Present address: Department of Ecology and Evolution, University of Chicago, Chicago, IL

[4]Present address: Center for Communicable Disease Dynamics and Department of Epidemiology, Harvard School of Public Health, Boston, MA

*Corresponding author: E-mail: hhchang@hsph.harvard.edu.

[†]These authors contributed equally to this work.

Abstract

Nearly 10% of the genes in the genome of *Drosophila melanogaster* are in nested structures, in which one gene is completely nested within the intron of another gene (nested and including gene, respectively). Even though the coding sequences and untranslated regions of these nested/including gene pairs do not overlap, their intimate structures and the possibility of shared regulatory sequences raise questions about the evolutionary forces governing the origination and subsequent functional and evolutionary impacts of these structures. In this study, we show that nested genes experience weaker evolutionary constraint, have faster rates of protein evolution, and are expressed in fewer tissues than other genes, while including genes show the opposite patterns. Surprisingly, despite completely overlapping with each other, nested and including genes are less likely to display correlated gene expression and biological function than the nearby yet nonoverlapping genes. Interestingly, significantly fewer nested genes are transcribed from the same strand as the including gene. We found that same-strand nested genes are more likely to be single-exon genes. In addition, same-strand including genes are less likely to have known lethal or sterile phenotypes than opposite-strand including genes only when the corresponding nested genes have introns. These results support our hypothesis that selection against potential erroneous mRNA splicing when nested and including genes are on the same strand plays an important role in the evolution of nested gene structures.

Key words: nested genes, overlapping genes, splicing, gene expression, gene organization.

The distribution of genes in the genome is not random. There are regions with few functional genes and regions where genes are densely packed. It has been known that the close proximity between genes can have significant functional consequences. Indeed, neighboring genes were shown to have correlated expression patterns in eukaryotes (including yeast [Cohen et al. 2000], *Caenorhabditis elegans* [Lercher et al. 2003], *Drosophila* [Boutanaev et al. 2002], *Arabidopsis thaliana* [Williams and Bowles 2004], and humans [Lercher et al. 2002; Trinklein et al. 2004]), as well as biological functions and/or signaling pathways (Elo et al. 2003; Lee and Sonnhammer 2003; Al-Shahrour et al. 2010). In extreme cases, the distance between neighboring genes is 0, and parts or all of their gene structures (exons, introns, or untranslated regions [UTRs]) overlap with each other (overlapping genes). These structures are commonly observed in eukaryotes (e.g., *C. elegans* [Chen and Stein 2006], *Drosophila* [Misra et al. 2002], and mammals [Veeramachaneni et al. 2004]).

An especially interesting class of overlapping genes is in which one gene is completely nested within an intron of another gene (nested and including gene, respectively [reviewed in Kumar 2009]). Even though the coding sequences of these nested/including gene pairs do not overlap, their intimate structures raise questions about the evolutionary forces governing the origination of nested gene structures and their subsequent functional and evolutionary impacts. We found that, in *Drosophila melanogaster,* approximately 16% of the genes (2,295 out of 14,072 genes) overlap with at least one other gene in exons, introns, or UTRs. Genes in nested structures account for 9.5% of the *D. melanogaster* genes (1,338 genes), which is more than *C. elegans* (2.7%, Chen and

Stein 2006) and human (2.73%, Yu et al. 2005). To examine the evolutionary and functional significance of nested gene structures in *D. melanogaster* while controlling for intrinsic attributes of genes in close proximity, we compared nested/including gene pairs to "control gene pairs," which have matching chromosomal distributions to that of nested/including gene pairs and are within 500 bp of each other but do not overlap (see Materials and Methods).

Mutational Input Is a Key Determinant of the Location of Nested Genes

Previous analysis showed that most nested gene structures in *Drosophila* originated through insertions or de novo origination of coding sequences in introns (Assis et al. 2008). Larger introns are larger targets for insertion or de novo mutations and should be more likely to harbor nested genes. Indeed, we found that the total intron lengths of including genes are significantly longer than control genes, even after excluding the sequence contributed by nested genes (median: 12,183 [including] and 308 [control]; Mann–Whitney U test (MWU) $P < 10^{-16}$). Including genes also have more introns than both nested genes and control genes (median: 7 [including], 2 [control], and 1 [nested]; MWU, $P < 10^{-16}$ for both comparisons). Focusing on including genes, introns with nested genes are significantly longer than introns without nested genes (median: 4,826 [with nested genes] and 138 [without nested genes]; MWU, $P < 10^{-16}$). Because long introns were found to be more evolutionarily conserved and suggested to more likely harbor functional sequences (Haddrill et al. 2005), this observation is unlikely due to larger introns being more tolerant of insertions. Moreover, the *D. melanogaster*–*D. simulans* divergence of the longest introns of including genes is smaller than that of other introns of including genes even after excluding nested genes (median: 0.071 [longest] and 0.082 [others]; MWU test, $P = 0.0012$), indicating that the observation of long introns being more evolutionarily conserved does not result from a fraction of nested genes in them. These results support that the mutational process is a key determinant of the location of nested genes.

Selection Plays an Important Role in the Maintenance and Functional Significance of Nested Gene Structures

Several hypotheses that potentially explain the selective pressures influencing the fixation of nested structures in the population, and their subsequent functional evolution, make specific predictions about the current expressional and functional correlations of nested and including genes. In addition to the common chromosomal environment that might have led to correlated expression of genes in proximity (reviewed in Hurst et al. 2004; Oliver and Misteli 2005), genes in nested structures might be selectively favored if their expression and/or biological functions are coregulated, resulting in even stronger positively correlated expression and/or biological functions

than neighboring genes. On the other hand, the proximity of nested and including genes may result in interference during transcription, leading to selection against spatially and temporally correlated expression of nested and including genes ("transcriptional interference" [Shearwin et al. 2005; Liao and Zhang 2008]). Still, the evolution of nested gene structures could be a nearly neutral process (Lynch and Conery 2003; Lynch 2006), and the expression and functional correlations between nested and including genes would be similar to those of genes in proximities.

Nested/including gene pairs are significantly positively correlated (estimated using Spearman rank ρ) in gene expression levels across tissues (FlyAtlas, Chintapalli et al. 2007, MWU, $P = 0.025$). This is also observed for control gene pairs (MWU, $P < 2 \times 10^{-16}$]. However, the correlations in expression of nested/including gene pairs are significantly weaker (Spearman rank ρ median 0.019 [nested/including gene pairs] vs. 0.174 [control gene pairs], MWU, $P = 8.6 \times 10^{-14}$, fig. 1) and less likely to be positive (52.74% [nested/including gene pairs] vs. 69.44% [control gene pairs]; Fisher's exact test [FET], $P = 4 \times 10^{-9}$) than control gene pairs. In fact, the correlations in expression of nested/including gene pairs are not different from two randomly chosen genes that are not adjacent but on the same chromosome ("random control gene pairs"; Spearman rank ρ median 0.019 [nested/including gene pairs] vs. 0.032 [random control gene pairs]; MWU, $P = 0.76$, fig. 1). Furthermore, we employed logistic regression and found that nested/including gene pairs are less likely than control gene pairs to have one gene (nested gene of nested/including gene pairs) to be expressed in the subset of tissues of another gene (including gene of nested/including gene pairs; $P = 0.05$; odds ratio $= 0.78$), to have the same highest expressed tissues ($P = 8 \times 10^{-11}$; odds ratio $= 0.25$), and to be associated with the same GO (Gene Ontology) categories ($P = 0.002, 0.001, 0.02$; odds ratios $= 0.14, 0.17, 0.16$ for biological process, molecular function, and cellular component, respectively). Yet, again, when we compared nested/including gene pairs with random control gene pairs, none of these three differences were significant. The correlations in expressional patterns and involvement in biological functions of nested/including genes pairs are significantly different from what have been observed for nearby nonoverlapping genes, suggesting that selection against transcriptional interference might have led to their expression in different tissues and involvement in different biological functions.

Paucity of Same-Strand Nested/Including Gene Pairs Might Result from Selection against Missplicing

Nested genes can be transcribed from the same strand as their including genes (same strand) or different strand from their including genes (opposite strand). The majority of nested genes (71.27%) were found to be on the opposite strand. This proportion is significantly different from the proportion
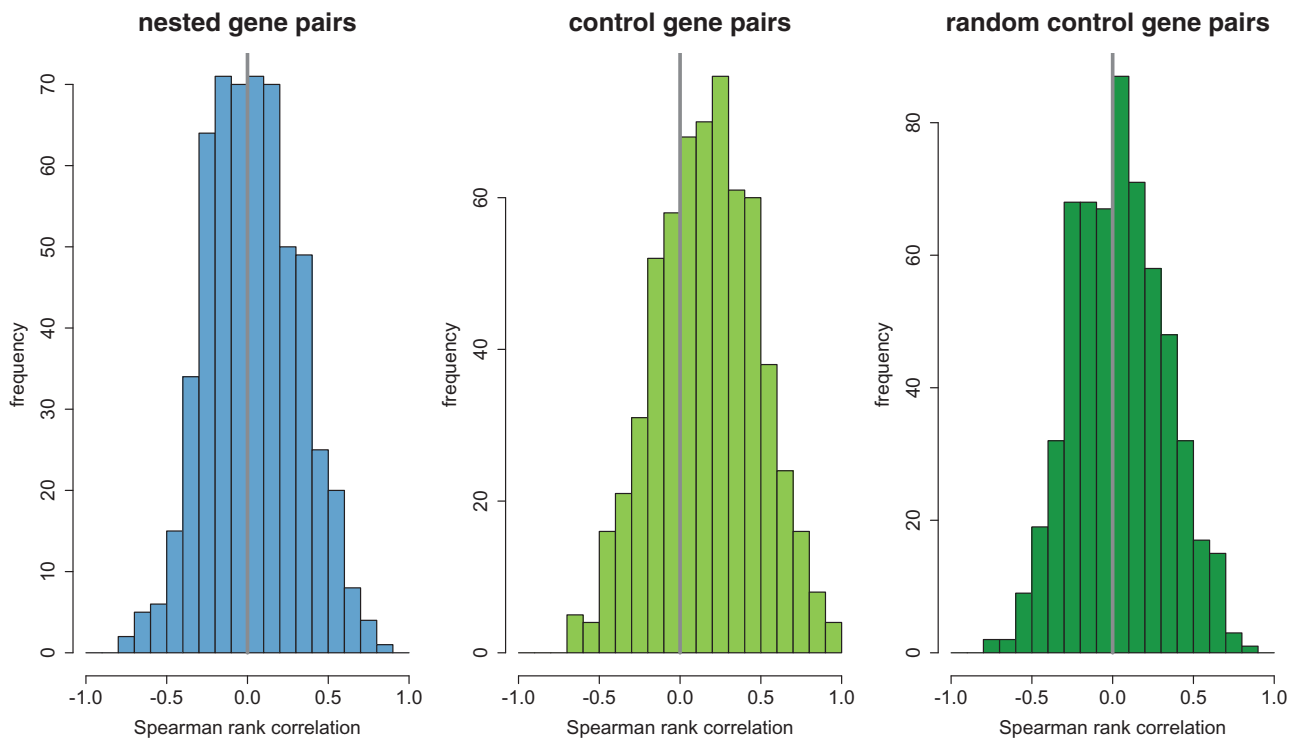
Fɪɢ. 1.—Distributions of Spearman ρ in gene expression for nested/including gene pairs and control gene pairs. Nested/including gene pairs are less positively correlated in their expression level across 20 tissues than control gene pairs, but have similar correlations in expression with nonadjacent pairs of genes on the same chromosome ("random control gene pairs").

of the control gene pairs (53.55%) and from the expected proportion if the orientations are random (50%; FET, $P < 10^{-16}$ for both comparisons). Although the strand biases of nested genes have been widely reported in different eukaryotes (63% of same-strand nested genes in human [Yu et al. 2005] and 88% in *C. elegans* [Chen and Stein 2006]), the biological cause of this bias has not been specifically discussed and tested on a genomic scale.

The paucity of same-strand nested gene structures may have resulted from the intrinsic strand biases of the mutational processes leading to nested gene structures. Alternatively, this may be due to differential selection on same-strand and opposite-strand nested genes. Several cases of genes, transposable elements, or endogeneous retroviruses that are nested within introns of another genes are known to cause aberrant splicing of the outer including genes (Horowitz and Berg 1995; Kaer et al. 2011; Maksakova et al. 2006). The missplicing of including genes was shown to be dependent on the presence of splice sites within the sequences of transposable elements or endogeneous virus (van de Lagemaat et al. 2006; Kaer et al. 2011). The splice sites of nested genes are more likely to interfere with splicing of including genes when the two genes are transcribed from the same strand. Consistent with this hypothesis, we found that same-strand nested genes are more likely to be single-exon genes (72.53%) than

opposite-strand nested genes (37.41%; FET, $P < 10^{-16}$). Focusing on nested genes that have more than one exon, same-strand nested genes still have fewer introns than opposite-strand nested genes (median: one intron (same-strand nested genes) vs. two introns [opposite-strand nested genes]; MWU, $P = 0.00013$). Our observation is not due to opposite-strand nested genes being longer than their same-strand counterparts because the coding sequence length is not statistically different between same-strand and opposite-strand nested genes (median: 817.5 [same strand] vs. 898 [opposite strand]; MWU, $P = 0.11$).

Seventy-three nested genes are young (less than 35 million years old [Clark et al. 2007; Zhang et al. 2010]) and originated through duplication of another gene (parental gene). The duplication process can be via either DNA or RNA intermediates. A characteristic of RNA-based duplication is that the new genes lose all introns that were originally present in their parental gene (reviewed in Kaessmann et al. 2009), and this process accounts for around 12.10% of duplicated genes in *Drosophila* (Zhang et al. 2010). Among the 73 duplicated nested genes, only 16.67% of opposite-strand nested duplicated genes originated through RNA-based duplication, while 42.11% of same-strand nested duplicated genes originated via RNA intermediates (FET, $P = 0.054$). This difference is marginally significant, likely due to the small sample size.

Table 1

Known Phenotypic Effects of Including Genes

		Lethal	Sterile	Viable	FET P Value	
					Lethal vs. Nonlethal[a]	Affected[b] vs. Viable
All including genes	Same strand	68	9	98	0.23	0.2
	Opposite strand	159	19	178		
Including genes with intron-containing nested genes	Same strand	13	2	35	0.037	0.027
	Opposite strand	91	11	113		

[a]Genes without known lethal phenotype (could have known sterile phenotype).
[b]Genes with known lethal or sterile phenotype.

Additionally, the decrease in intron number of duplicated nested genes, when compared with their respective parental genes, is significantly larger for same-strand nested duplicated genes than opposite-strand nested duplicated genes (median: one intron difference [same-strand nested genes] vs. zero intron difference [opposite-strand nested genes]; MWU, $P = 0.028$). Note that this difference is not due to the variation in intron numbers of the parental genes of same-strand and opposite-strand nested genes, which is not significantly different (MWU, $P = 0.41$).

If missplicing is indeed more likely to happen when including genes and nested genes are on the same strand than when they are on opposite strands, we expect that same-strand including genes are less likely to be essential for the fitness of flies. In extreme cases, we expect that loss of function or expression knockdown by RNA interference (RNAi) of same-strand including genes is less likely to be associated with lethal phenotypes. When considering all same-strand and opposite-strand including genes, there is no significant difference in the proportion of genes having known lethal phenotypes (38.85% [same strand] vs. 44.66% [opposite strand]; table 1). Yet, when we only considered including genes whose nested genes have introns (and therefore are more likely to cause missplicing), same-strand including genes are significantly less likely to have known lethal phenotypes (26.0% [same strand] vs. 42.33% [opposite strand]; table 1). The result is strengthened if we consider both lethal and sterile phenotypes (30.00% [same strand] vs. 47.44% [opposite strand]; table 1). It is worth noting that the genetic disturbance (null mutant or expression knockdown) we considered here is extreme, and it is likely that, when considering more subtle influences on fitness, the difference between same-strand and opposite-strand including genes will be more significant and should be more general. Overall, our observations that same-strand nested genes contain fewer introns and that same-strand including genes have a lower probability of being associated with lethal and sterile phenotypes suggest that the paucity of same-strand nested/including gene pairs could be attributable to purifying selection against missplicing when nested genes are transcribed from the same strand.

Nested Genes Evolve Faster, Are More Narrowly Expressed, and Are Enriched with Testis-Related Functions While Including Genes Show the Opposite Patterns

To test whether genes in nested structures show different patterns of evolution, we examined the site frequency spectrum of coding variants (using Tajima's D [Tajima 1989]), relative rates of protein evolution (d_N/d_S, [Yang 2007]), and proportion of amino acid substitutions fixed by positive selection (α, [Smith and Eyre-Walker 2002]) of including genes, nested genes and control genes, and classified genes into those that are present in all 12 *Drosophila* species (i.e., genes older than 35 million years; Clark et al. 2007) or not (Zhang et al. 2010) (table 2). Including genes have more negative Tajima's D, lower d_N/d_S, and are more likely to be conserved across the *Drosophila* species than either nested genes or control genes, suggesting they are under stronger purifying selection. On the other hand, nested genes, while not differing in Tajima's D from control genes, have larger d_N/d_S and α, and tend to be younger than both including genes and control genes. We did not detect any significant difference between same- and opposite-strand including genes or nested genes in these analyses.

We also found that nested and including genes have unusual gene expression patterns. Nested genes are expressed in significantly fewer tissues (have narrower breadth of expression) than either including genes or control genes (table 2). They also have significantly higher expression specificity (see Materials and Methods) than either including or control genes (MWU, $P < 10^{-12}$ for both comparisons; fig. 2). While same- and opposite-strand nested genes do not differ in their breadth of expression (MWU, $P = 0.15$), same-strand nested genes have significantly higher expression specificity than opposite-strand nested genes (0.95 [same-strand] vs. 0.93 [opposite-strand]; MWU, $P = 0.009$). The composition of tissues where genes have their highest expression is also significantly different between including genes, nested genes, and control genes (chi-square test, $P < 10^{-16}$ for all comparisons; fig. 3). This composition is not different between same- and opposite-strand including genes but significantly different between same- and opposite-strand nested genes (chi-square test,

Table 2

Evolutionary Properties and Expression Patterns of Nested, Including, and Control Genes

	Median			MWU Test P Value		
	Including	Nested	Control	Including vs. Nested	Including vs. Control	Nested vs. Control
Tajima's D	-2.76	-1.77	-1.87	$<10^{-8}$	$<10^{-8}$	>0.05
d_N/d_S	0.042	0.107	0.073	$<10^{-8}$	$<10^{-8}$	$<10^{-8}$
α	0.251	0.435	0.343	0.005	0.275	0.035
Breadth of expression (number of tissues)	18	4	19	$<10^{-16}$	0.363	$<10^{-16}$

|	Proportion (%)			FET P value		
	Including	Nested	Control	Including vs. Nested	Including vs. Control	Control vs. Nested
Conserved across 12 *Drosophila* species	99.05	88.13	91.24	$<10^{-16}$	$<10^{-16}$	0.027
Highest expression in brain	29.09	5.21	9.44	$<10^{-16}$	$<10^{-16}$	0.003
Highest expression in testis	6.43	43.91	13.52	$<10^{-16}$	1.45×10^{-6}	$<10^{-16}$
Highest expression in ovary	13.78	5.36	23.94	1.3×10^{-7}	9.06×10^{-8}	$<10^{-16}$
Young duplicate genes	0.9	8.4	7	$<10^{-12}$	5.2×10^{-16}	0.02
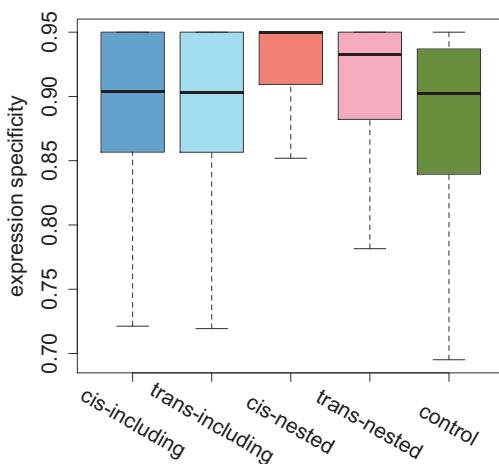
Fig. 2.—Expression specificity of genes in nested structures and control genes. Boxplots for the expression specificity of including genes, nested genes, and control genes. The expression specificity is highest for same-strand nested genes followed by opposite-strand nested genes, both of which are significantly higher than either including genes or control genes.

$P = 0.024$; fig. 3). Including genes are more enriched with genes having their highest expression in brain than either nested genes or control genes (table 2). In contrast, nested genes are significantly enriched with genes having highest expression in testis but are deficient for genes having highest expression in ovaries (table 2). The enrichment of high testis expression is especially strong for same-strand nested genes (58.46% [same strand] vs. 38.18% [opposite strand]; FET, $P = 1.67 \times 10^{-6}$).

Consistent with previous finding that the majority of nested gene structures originated through insertion of DNA sequences into introns of including genes via gene duplications (Assis et al. 2008), we observed significantly larger proportion of nested genes that were previously identified as young duplicated genes (Zhang et al. 2010) than either including genes or control genes (table 2). Young duplicated genes tend to evolve rapidly (Chen et al. 2010), which could have led to the observed exceptional evolutionary properties of nested genes. On the other hand, the two interesting properties of nested genes—narrow expression (Larracuente et al. 2008) and enrichment of highest expression in testis (reviewed in Swanson and Vacquier 2002)—are widely known to be correlated with rapid protein evolution. To test whether the unusual evolutionary and expression properties of nested genes are due to the larger proportion of duplicate genes, we compared nested genes with a set of control genes that have the same proportion of young duplicated genes ("duplication control genes," see Materials and Methods). Nested genes still show faster rates of protein evolution (d_N/d_S, MWU, $P < 10^{-9}$), have greater α (MWU, $P = 0.0021$], are expressed in fewer tissues (MWU, $P < 10^{-16}$), have higher expression specificities (MWU, $P < 10^{-16}$), and are enriched with genes having highest expression in testis (FET, $P < 10^{-16}$). These results indicate that the observed patterns could not be simply explained by the higher proportion of duplicate genes. On the contrary, when using another set of control genes that have the same expression patterns as nested genes ("expression control genes," see Materials and Methods), nested genes are not significantly different from control genes with respect to d_N/d_S, α, or gene age (MWU, $P > 0.05$ for all comparisons). Accordingly, the evolutionary properties of nested genes might have been the "by-product" of their expressional
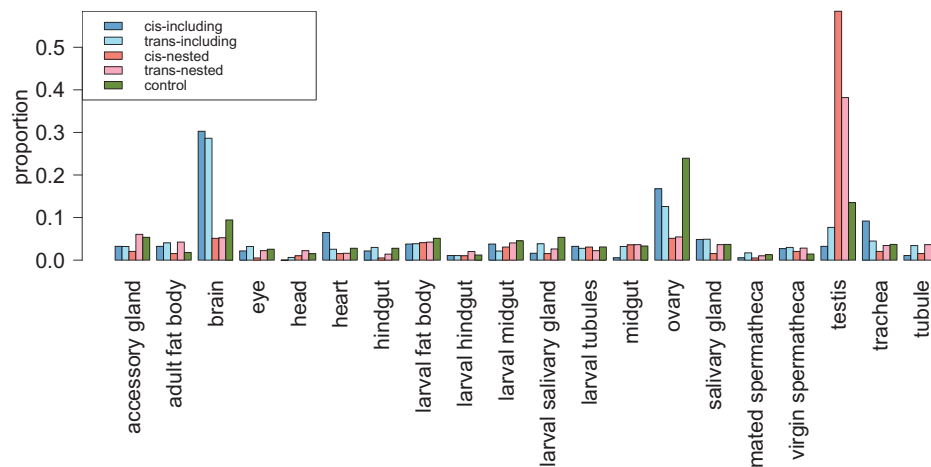
FIG. 3.—The distributions of tissues where genes have their highest expression. Nested genes, especially same-strand nested genes, are enriched with genes having their highest expression level in testis when compared with both including and control genes. On the contrary, including genes are enriched with genes having their highest expression in brain.

attributes. However, selection to decouple the functions of nested genes from those of including genes due to their nested structures could have led to the observed narrow expression of nested genes and could be the ultimate cause for the evolutionary properties of nested genes.

While including genes are slowly evolving, highly conserved, broadly expressed, and enriched with genes having their highest expression in brain, nested genes are the opposite: fast evolving, narrowly expressed, and enriched with genes having their highest expression in testis. Thus, positive selection for coregulation in gene expression and biological function, which might have driven the evolution of gene clusters (reviewed in Hurst et al. 2004), is unlikely to apply to the fixation of nested gene structures. The fixation of nested gene structures, similarly to evolution of other complex genomic organizations (Lynch and Conery 2003; Lynch 2006), could have been a nearly neutral process. However, we have evidence supporting the role of natural selection in shaping the relative orientations and functional importance of nested gene structures. We showed that nested/including gene pairs are less likely to be transcribed from the same strand and that same-strand nested genes are more likely to be single-exon genes and have fewer exons if they are multiexon genes. Together with the finding that including genes with same-strand nested genes that contain introns are less likely to be essential for fitness of flies, our results support that selection against missplicing events of same-strand nested/including gene pairs leads to this bias. In addition, the correlations in expressions and biological functions of nested/including gene pairs are lower than those of nearby gene pairs but similar to any two random genes of the same chromosome. This is consistent with the hypothesis that selection against transcriptional interference plays an important role in shaping the

functional significance and indirectly affects evolutionary properties of nested gene structures. In sum, despite the proximity of nested and including genes, we found that they are nowhere similar to each other in terms of evolutionary properties, expressional patterns, and biological functions, and selection against the potential deleterious impacts caused by their close proximity might have been the main force governing their evolution.

Materials and Methods

We used *D. melanogaster* genome annotation version 5.47 and only considered coding transcripts that are annotated as "strongly supported" by FlyBase. For genes that had more than one isoform in nested gene structures, we considered the isoform with the longest coding sequence. We used FlyAtlas Expression data (www.flyatlas.org [last accessed October 17, 2013], Chintapalli et al. 2007), which used four microarrays to measure gene expression for each of the 20 tissues of *D. melanogaster* at various developmental stages. Genes were considered expressed in a tissue if annotated as "presence" by FlyAtlas for at least three of the four microarrays, and highest expressed tissues were determined by using mean expression levels. Expression specificity, an index between 0 and 1, was calculated as described in the previous paper (Yanai et al. 2005). Broadly expressed genes with similar expression level in all tissues have low indices while tissue-specific genes have high indices. We used the population genomic sequences from Drosophila Population Genomic Project (DPGP, www.dpgp.org [last accessed October 17, 2013]; Langley et al. 2012) and the multispecies alignments (including *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. erecta*) as described in Langley et al. (2012) to perform evolutionary genetic analyses. d_N/d_S on the branch leading

to *D. melanogaster* was estimated using *D. melanogaster*, *D. simulans*, and *D. yakuba* alleles, and *PAML* (phylogenetic analysis by maximum likelihood) (HKY85 nucleotide substitution model [Hasegawa et al. 1985] and free-ratio branch model [model = 1]). α (McDonald and Kreitman 1991; Smith and Eyre-Walker 2002) was estimated using *D. melanogaster* within-species polymorphism (using both African and North American alleles of DPGP, total 44 alleles) and *D. simulans* as an outgroup. Tajima's *D* was calculated using North American *D. melanogaster* population of DPGP, which has a much larger sample size (37) than the African population (7). Gene sizes include the length of coding sequences, intron, and both UTRs. We batch downloaded phenotype data associated with nested genes from FlyBase and classified a gene to be associated with known lethal/sterile phenotype if at least one mutation or one expression knockdown experiment (using RNAi) was reported to be lethal/sterile. Genes that have both reported lethal and sterile phenotypes are denoted as only lethal phenotype.

Control gene pairs are gene pairs less than 500 bp apart and were chosen randomly. The chromosomal distributions were matched: the number of control gene pairs on each chromosome was matched to the number of nested/including gene pairs on the same chromosome. We also used two other sets of control gene pairs: 1) gene pairs that are not adjacent to each other but are on the same chromosome ("random control gene pairs"); 2) gene pairs that are less than 500 bp apart and have the same chromosomal distributions and the same proportion of same/opposite-strand as nested/including gene pairs ("same/opposite control gene pairs"). Our observed correlations in expressional patterns and functional categories of nested/including gene pairs hold when we compared them with same/opposite control gene pairs, but differ when comparing with random control gene pairs (see details in main text). Other comparisons of evolutionary properties and expression patterns between nested/including genes and control genes from different sets of control gene pairs are consistent.

In addition to further tease apart the main evolutionary force that might have led to some of our observations, we generated additional control gene sets that match either the proportion of DNA/RNA-duplicated genes of nested genes (duplication control genes) or gene expression of nested and including genes (expression control genes). Duplication control genes were matched to have the same proportions of DNA and RNA duplicate genes and chromosomal distributions as nested genes. Expression control genes were matched to have the same proportion of expression in testis and brain, the same chromosomal distributions, and similar expression breath and specificity (within 0.25 standard deviations) with nested genes and including genes individually.

We used FET when examining a relationship between two categorical variables and MWU for noncategorical variable between two conditions. We employed logistic regression when studying how a binary outcome variable changes with input variables. The odds ratio was obtained by raising the natural exponent e to the power of the logistic coefficient. Because nested genes have significantly narrower breadth of expression (see earlier), and, accordingly, nested genes are more likely to be expressed in a subset of tissues of including genes by chance. We thus include the number of tissues being expressed as a covariate in the analysis. Otherwise, the only covariate is the binary variable that represents whether it is nested/including gene pair (=1) or control gene pair (=0), and the response variable is the property that is examined. All statistical analyses were performed using R (http://www.R-project.org [last accessed October 17, 2013], R Development Core Team 2006).

Acknowledgments

Literature Cited

Al-Shahrour F, et al. 2010. Selection upon genome architecture: conservation of functional neighborhoods with changing genes. PLoS Comput Biol. 6:e1000953.

Assis R, Kondrashov AS, Koonin EV, Kondrashov FA. 2008. Nested genes and increasing organizational complexity of metazoan genomes. Trends Genet. 24:475–478.

Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI. 2002. Large clusters of co-expressed genes in the *Drosophila* genome. Nature 420: 666–669.

Chen N, Stein LD. 2006. Conservation and functional significance of gene topology in the genome of *Caenorhabditis elegans*. Genome Res. 16: 606–617.

Chen S, Zhang YE, Long M. 2010. New genes in *Drosophila* quickly become essential. Science 330:1682–1685.

Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet. 39: 715–720.

Clark AG, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218.

Cohen BA, Mitra RD, Hughes JD, Church GM. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet. 26:183–186.

Elo A, Lyznik A, Gonzalez DO, Kachman SD, Mackenzie SA. 2003. Nuclear genes that encode mitochondrial proteins for DNA and RNA metabolism are clustered in the *Arabidopsis* genome. Plant Cell 15: 1619–1631.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. Genome Biol. 6:R67.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22: 160–174.

Horowitz H, Berg CA. 1995. Aberrant splicing and transcription termination caused by P element insertion into the intron of a *Drosophila* gene. Genetics 139:327–335.

Hurst LD, Pál C, Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 5:299–310.

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:e26099.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet. 10:19–31.

Kumar A. 2009. An overview of nested genes in eukaryotic genomes. Eukaryot Cell. 8:1321–1329.

Langley CH, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics 192:533–598.

Larracuente AM, et al. 2008. Evolution of protein-coding genes in *Drosophila*. Trends Genet. 24:114–123.

Lee JM, Sonnhammer ELL. 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13:875–882.

Lercher MJ, Blumenthal T, Hurst LD. 2003. Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. Genome Res. 13:238–243.

Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet. 31:180–183.

Liao BY, Zhang J. 2008. Coexpression of linked genes in mammalian genomes is generally disadvantageous. Mol Biol Evol. 25:1555–1565.

Lynch M. 2006. The origins of eukaryotic gene structure. Mol Biol Evol. 23: 450–468.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401–1404.

Maksakova IA, et al. 2006. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet. 2:e2.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351:652–654.

Misra S, et al. 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol. 3:0083.1–83.22.

Oliver B, Misteli T. 2005. A non-random walk through the genome. Genome Biol. 6:214.

Shearwin KE, Callen BP, Egan JB. 2005. Transcriptional interference–a crash course. Trends Genet. 21:339–345.

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. Nature 415:1022–1024.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet. 3:137–144.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Trinklein ND, et al. 2004. An abundance of bidirectional promoters in the human genome. Genome Res. 14:62–66.

van de Lagemaat LN, Medstrand P, Mager DL. 2006. Multiple effects govern endogenous retrovirus survival patterns in human gene introns. Genome Biol. 7:R86.

Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I. 2004. Mammalian overlapping genes: the comparative perspective. Genome Res. 14:280–286.

Williams EJB, Bowles DJ. 2004. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res. 14:1060–1067.

Yanai I, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinforma Oxf Engl. 21:650–659.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Yu P, Ma D, Xu M. 2005. Nested genes in the human genome. Genomics 86:414–422.

Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2010. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. Genome Res. 20:1526–1533.

Associate editor: Esther Betran