Published in final edited form as:
Evol Dev. 2011 ; 13(1): 58–71. doi:10.1111/j.1525-142X.2010.00456.x.

Differential Selection within the Drosophila Retinal Determination Network and Evidence for Functional Divergence between Paralog Pairs

Rhea R. Datta, **Tami Cruickshank**, and **Justin P. Kumar***
Department of Biology, Indiana University Bloomington, IN 47405

Abstract

The retinal determination (RD) network in *Drosophila* comprises fourteen known nuclear proteins that include DNA binding proteins, transcriptional co-activators, kinases and phosphatases. The composition of the network varies considerably throughout the animal kingdom, with the network in several basal insects having fewer members and with vertebrates having potentially significantly higher numbers of retinal determination genes. One important contributing factor for the variation in gene number within the network is gene duplication. For example, ten members of the RD network in *Drosophila* are derived from duplication events. Here we present an analysis of the coding regions of the five pairs of duplicate genes from within the retinal determination network of several different *Drosophila* species. We demonstrate that there is differential selection across the coding regions of all RD genes. Additionally, some of the most significant differences in ratios of non-silent to silent site substitutions (d_N/d_S) between paralog pairs are found within regions that have no ascribed function. Previous structure/function analyses of several duplicate genes have identified areas within one gene that contain novel activities when compared to its paralog. The evolutionary analysis presented here identifies these same areas in the paralogs as being under high levels of relaxed selection. We suggest that sequence divergence between paralogs and selection signatures can be used as a reasonable predictor of functional changes in rapidly evolving motifs.

Keywords

Drosophila; retinal determination; phylogeny; gene duplication; gene regulatory network

Introduction

Gene duplications can have a profound impact on signal transduction pathways and gene regulatory networks. Upon duplication a number of evolutionary paths can be taken by either of the two paralog genes. In one scenario both genes remain and are functionally redundant (Gibert, 2002; Hughes, 1994; Hurley et al., 2005; Krakauer and Nowak, 1999; Ohta, 1989; Wagner, 1996) while at the other extreme one of the two paralogs becomes a pseudogene and is subsequently lost (Balakirev and Ayala, 2003; Force et al., 1999; Harrison et al., 2003; Lynch and Conery, 2003; Ohta, 1989; Vanin, 1985).Wedged between these extremes are two outcomes that are more relevant to understanding the evolution of regulatory circuits: neo-functionalization (where one copy acquires a completely novel function) and sub-functionalization (where the function of the ancestral gene is divided amongst the two daughter genes; (Lynch and Conery, 2000). The latter two situations are

*Correspondence should be addressed to: Justin P. Kumar jkumar@indiana.edu.

particularly relevant for the functioning of developmental systems and thus many duplicate genes have become the objects of extensive structure/function studies.

Traditional methods for comparing the functions of paralog genes involve molecular dissections of the two proteins followed by the use of these modified molecules in one or more functional assays such as rescue and/or forced expression tests. Based on the phenotypic results of these studies, it is often possible to determine if functional differences between the paralogs have been acquired and to map the putative new functional motifs. This information is important for understanding how a gene regulatory network or a signal transduction cascade has evolved and for understanding how individual proteins function during development. However, differences between paralogs are not *a priori* apparent, thus most structure function studies are conducted using laborious brute force approaches. Additionally, the mechanisms underlying functional divergence amongst genes are difficult to characterize without cross-species analysis for which tools are limited despite huge strides in research over the past decade. Gene duplications, which often are a large part of developmental networks, provide nice internal controls for rates of evolution and changes in gene structure as paralogs have diverged for the same amount of time. Previous studies support theoretical models of differential subfunctionalization, but data from additional developmental processes are required to identify regions of change within paralogs (Dermitzakis and Clark, 2001; Lynch and Force, 2000). Here we have attempted to devise a new strategy that uses selection signatures across coding regions to identify new functional domains or motifs in paralog pairs. The results presented in this paper suggest that a sequence based analysis can be used to guide structure/function studies and this allows for more targeted molecular dissections of proteins.

We have examined the levels of selection across full-length genes and functional domains along the coding regions of the highly characterized *Drosophila* retinal determination network genes as part of an effort to see if the areas with the highest rates of differential selection coincide with regions that have been identified (from structural studies) as having acquired new functional domains. The retinal determination network was chosen as the subject of our analysis since ten of the fourteen known members (71.4%) of this network are the products of gene duplication events (Kumar, 2009a) and since such events, which are some of the most important factors in evolution (Ohno, 1970), also greatly influences the development of gene regulatory networks (Amoutzias et al., 2004; Chen et al., 2007; Gardiner et al., 2008; Gibert, 2002; Gu et al., 2004; Hughes and Friedman, 2005; Rudel and Sommer, 2003; Shimeld, 1999; Teichmann and Babu, 2004; Wagner, 1996). As currently understood, the retinal determination network in *Drosophila* includes fourteen genes that code for DNA binding proteins and transcriptional co-activators as well as protein kinases and phosphatases (Kumar, 2009a). Within this set are five pairs of duplicate genes: the Pax6 genes *eyeless* and *twin of eyeless* (*ey, toy*: Czerny et al., 1999; Quiring et al., 1994), the Pax6(5a) genes *eyegone* and *twin of eyegone* (*eyg, toe*: Aldaz et al., 2003; Jun et al., 1998), the Six family members *sine oculis* and *optix* (*so, optix*: Cheyette et al., 1994; Seimiya and Gehring, 2000; Serikaku and O'Tousa, 1994), the Tsh class genes *teashirt* and *tiptop* (*tsh, tio*: Laugier et al., 2005; Pan and Rubin, 1998) as well as the pipsqueak genes *distal antenna* and *distal antenna related* (*dan, danr*: Curtiss et al., 2007). The remainder of the network is made up of single members of the Eya, Dach, Meis and Nlk gene families and are represented by *eyes absent* (*eya*: Bonini et al., 1993), *dachshund* (*dac*: Mardon et al., 1994), *homothorax* (*hth*: Pai et al., 1998) and *nemo* (*nmo*: Braid and Verheyen, 2008; Choi and Benzer, 1994). With some exceptions each gene family is required for retinal development in all seeing animals examined so far including mice and humans.

We also chose the retinal determination network as it is one of the most extensively studied gene regulatory networks in both invertebrate and vertebrate systems. Not only are the gene

families and signaling pathways that specify the compound eye highly conserved across species but so also are the developmental defects that are associated with mutations in these genes (Callaerts et al., 1997; Donner and Maas, 2004; Gehring, 1996; Gehring and Ikeo, 1999; Hanson, 2001; Jean et al., 1998; Kumar, 2001; Kumar, 2009a; Kumar, 2009b; Treisman, 1999; Wawersik and Maas, 2000). In *Drosophila* mutations within most network members result in severe reductions in eye development (Bonini et al., 1993; Cheyette et al., 1994; Curtiss et al., 2007; Jun et al., 1998; Mardon et al., 1994; Quiring et al., 1994; Serikaku and O'Tousa, 1994). Conversely, forced expression of these genes can coax certain cell populations within non-retinal tissues into adopting a retinal fate (Bonini et al., 1997; Braid and Verheyen, 2008; Curtiss et al., 2007; Czerny et al., 1999; Halder et al., 1995; Salzer and Kumar, 2010; Seimiya and Gehring, 2000; Shen and Mardon, 1997; Weasner et al., 2007). These phenotypes place members of the retinal determination network at the highest levels of the eye specification hierarchy. As loss-of-function phenotypes of several mouse models and human retinal disorders are very similar to those seen in *Drosophila* there is a considerable interest in understanding not only how the network functions as a unit but also how individual genes acquire new and novel functions.

In this paper we examine the selection signatures (defined by the varying levels of selection across the gene) along the coding regions of the RD genes, and have compared selection on paralog pairs with well known functions in an attempt to ascertain whether changes in function can be attributed to differential selection on the protein. In doing so, we have attempted to devise a new strategy that uses selection signatures across coding regions to identify new functional domains or motifs in paralog pairs. We have used the fully sequenced genomes of ten *Drosophila* species to identify putative orthologs of factors that are known to act during eye specification in *Drosophila melanogaster*. Using phylogenetic analyses on the aligned sequences we have measured the amount of divergence across different *Drosophila* species for each member of the cascade in order to determine how the genes in the cascade may be diverging as a whole within the *Drosophilid* lineage. Using the ratio of non-synonymous nucleotide substitutions (d_N) to synonymous substitutions (d_S), we have measured the rate of substitutions at non-silent sites (which are under selection) to silent sites (which are presumed neutral). We have used d_N/d_S ratios for each full-length gene as well as conserved domains with well-known functions such as the Homeo, Paired, Pipsqueak and Zn finger DNA binding motifs and the SIX protein-protein interaction motif. We also considered the non-conserved regions within each gene that have as of yet no ascribed function. These measurements have allowed us to determine the substitution rates and evolutionary constraints on various regions of each paralog pair. We have also used the *D. melanogaster* and *D. simulans* population data sets to confirm selection signatures. Our findings show that differential selection on the coding regions of paralog pairs correlate with empirical evidence of functional divergence of duplicates in the network, and that non-conserved domains are under more relaxed selection and are likely to gain new functions. Additionally, the data indicate that there is a disparity in the selection pressure across the non-conserved regions between paralogs. We propose that utilizing this approach will aid in the identification of specific regions that may be gaining new functions.

Materials and Methods

Gene and Domain Selection

Retinal determination genes from *D. melanogaster* were used as a reference set as input for tblastn searches against the other sequenced *Drosophila* genomes. We have selected the fourteen genes that comprise the traditional retinal determination cascade (eyeless, twin of eyeless, eyegone, twin of eyegone, sine oculis, optix, eyes absent, dachshund, homothorax, teashirt, tiptop, distal antenna, distal antenna related and *nemo*). We also included *DSix4* as it represents the sister gene of *sine oculis* and *optix*. The accession numbers for the *D.*

melanogaster sequences are CG1464, CG11186, CG10488, CG10704, CG11121, CG18455, CG3871, CG1374, CG12630, CG11849, CG13651, CG9554, CG4952, CG17117, and CG7892.

Using Flybase BLAST (http://flybase.bio.indiana.edu/blast/) and Gbrowse (http://flybase.bio.indiana.edu/cgi-bin/gbrowse/) the putative orthologs to the *D. melanogaster* retinal determination genes were identified in: *D. sechellia* (acc #s GM26810, GM13021, GM24656, GM24657, GM20951, GM21001, GM22186, GM16144, GM16153, GM17814, GM17808, GM14165, GM17175, GM26167, GM25010), *D. yakuba* (acc #s GE14559, GE14563, GE20121, GE20122, GE19124, GE19170, GE22381, GE12928, GE12934, GE23568, GE23564, GE18433, GE12814, GE24684, GE21651), *D. erecta* (acc #s GG16399, GG16402, GG13831, GG13832, GG23277, GG23326, GG13283, GG21347, GG21353, GG11370, GG11367, GG23613, GG20126, GG17283, GG14461), *D. ananassae* (acc #s GF22818, GF21877, GF24979, GF24980, GF12955, GF13656, GF24111, GF21464, GF21514, GF17896, GF17893, GF15700, GF15881, GF17520, GF23805), *D. persimilis* (acc #s GL18183, GL17563, GL24956, GL11165, GL11483, GL12751, GL25653, GL25663, GL21909, GL21906, GL26097, GL18480, GL23827, GL18049), *D. willistoni* (acc #s GK13702, GK13683, GK12628, GK12629, GK21318, GK23166, GK12359, GK18714, GK18717, GK13892, GK13890, GK14797, GK24323, GK22586, GK20588), *D. mojavensis* (acc #s GI14081, GI14042, GI12327, GI12329, GI18759, GI19347, GI11887, GI17699, GI17708, GI23731, GI23730, GI18181, GI13824, GI23559, GI13148), *D. virilis* (acc #s GJ15657, GJ13300, GJ13265, GJ13266, GJ21783, GJ22125, GJ13766, GJ11344, GJ11452, GJ23602, GJ23601, GJ14644, GJ17453, GJ23538, GJ13897), and *D. grimshawi* (acc #s GH24002, GH23963, GH16064, GH21164, GH21899, GH14784, GH22172, GH10455, GH18053, GH18050, GH13031, GH25029, GH19106, GH14986).

Using tblastn searches in NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi/; (Gertz et al., 2006) putative homologs of duplicate genes were identified from the following databases: *Anopheles gambiae* (str. PEST; release (3/22/2002); *Apis mellifera* (DH4; release (03/01/2006); *Tribolium castaneum* (GA2; release (8/17/2005; acc #s EU169112, NM001114345, XM967074, XM963647). Only coding regions were used in the analysis. The nucleotide sequences for each gene and their corresponding homologs in the other species were aligned using ClustalW (Thompson et al., 1994).

In addition to our analysis of full-length coding sequences we have examined the divergence and evolutionary constraints that have been placed on individual functional domains including both DNA binding and protein-protein interaction domains. The putative functional domains for each gene were annotated in the other *Drosophila* species based on alignment with previously defined functional domains in *D. melanogaster* (Aldaz et al., 2003; Cheyette et al., 1994; Clark et al., 2007; Curtiss et al., 2007; Czerny et al., 1999; Fasano et al., 1991; Jun et al., 1998; Laugier et al., 2005; Quiring et al., 1994; Seimiya and Gehring, 2000; Serikaku and O'Tousa, 1994). Several non-conserved regions were also included in the analysis due to the presence of experimentally verified transcriptional activation and/or repressor activity. An annotation of each protein can be found in Figure 1.

Phylogenetic Tree Construction, Evolutionary Distance Calculation, Substitution Rates

Neighbor-joining trees were generated using the Kimura 2-parameter model in MEGA v4 (Tamura et al., 2007), and although support for the majority of nodes was high, the topology of all trees was also verified using maximum likelihood in Paup* using the GTR+I+G model (Wilgenbusch and Swofford, 2003). Only minor differences in topology were found, particularly in the placement of *D. ananassae* in trees for *dan/danr* and *ey/toy*. All of the sequences used were full length with gaps and support for internal nodes were determined with 1000 bootstrap replicates. We calculated pair wise divergence between species and

total divergence across all species for each gene. In order to determine the evolutionary distance between different species for a particular gene, a distance-based tree was generated using the *Drosophilid* nucleotide sequences. Branch lengths were determined in MEGA using the red flour beetle, *Tribolium castaneum*, as an out-group. The ratio of non-synonymous to synonymous substitutions (d_N/d_S) was calculated for full-length genes, functionally conserved domains and non-conserved regions in the melanogaster group only (where the synonymous sites are not saturated). The Kumar method (which differentially corrects for multiple substitutions at different sites) was used with pair-wise deletions for missing sites with 1000 bootstrap replicates to calculate standard errors and 95% confidence intervals (Nei, 2000). Pair-wise deletions were used due to the short length of the specific domains. We were also interested in increasing resolution of our d_N/d_S estimates within known functional domains. In particular, the C terminals of Ey/Toy and Eyg/Toe. Therefore, we measured d_N/d_S (according to the methods described above) in non-overlapping 100bp intervals across the C terminal domains of ey/toy and eyg/toe (Supplemental Figure 1). We set a stringent cutoff of $d_N/d_S>1$ as an indication of recurrent positive selection. In our analysis $d_N/d_S=1$ meant neutral evolution, and $d_N/d_S<1$ was an indicator of purifying selection. A comparison of domains under various levels of purifying selection allowed us to distinguish between areas with higher numbers non-silent substitutions than others. These regions are described as being under relaxed constraint compared to those with fewer non-synonymous substitutions.

We also assessed selection using population sequence data. Specifically, we used genome sequence from 39 inbred lines of *D. melanogaster* from the Drosophila Population Genomics Project. For each gene in our analysis, we obtained sequence directly using coordinates from FlyBase, consistent with *D. melanogaster* Reference Version 4. For McDonald-Kreitman tests, we used *D. sechellia* and *D. yakuba* as our out-group species and counted synonymous and non-synonymous polymorphic sites and fixed differences. Significance levels were assessed using Fisher's exact tests. Combined with our estimates of d_N/d_S, this analysis allowed us to look for a signature of positive and negative selection on each gene in our sample. Population sequence analysis for *eyeless* and *twin of eyeless* was not conducted due to reduced coverage on the fourth chromosome.

In order to identify individual amino acid substitutions that could potentially serve as distinguishing markers between two proteins that are encoded by duplicate genes we compared the amino acid composition of the DNA binding and protein-protein interaction domains of the Ey/Toy, Eyg/Toe, Tsh/Tio, Dan/Danr and So/Optix/DSix4 protein pairs from ten *Drosophila* species (Table 2; Supplementary Figure 2). For each protein domain we have identified clade-specific amino acids, which may serve as marks for functionally distinguishing one protein from another. In order to be considered clade specific, a residue must fulfill two criteria. First, the orthologous position within two sister proteins (products of a duplication event) must be occupied by different amino acids, and second, this difference must be maintained in all ten *Drosophila* species. Second, we have also identified substitution events that have taken place relatively recently and are thus confined to a subset of *Drosophila* species.

Results

Evolutionary Constraints on the RD Network across the Drosophilidae

We were first interested in determining how the RD network as a whole is evolving across different species of the *Drosophila*. We identified the putative orthologs of the *D. melanogaster* network genes from the genome sequences of ten additional *Drosophila* species, one mosquito species (*Anopheles gambiae)*, one flour beetle species (*Tribolium castaneum)*, and the honeybee (*Apis mellifera*; Fig. 1A-C, see Materials and Methods). Each

gene encodes a DNA binding protein, with the exceptions of *eya* and *nemo*, which code for a transcriptional co-activator/protein tyrosine phosphatase and a serine-threonine kinase respectively (Kumar, 2009a). These factors are organized into a complicated regulatory system where at least one gene from each gene family has been conserved in organisms ranging from insects to vertebrates (Kozmik et al., 2007; Kumar, 2009a; Silver and Rebay, 2005). In order to examine the rates of evolution and the degree of sequence divergence we constructed distance-based trees for each gene family and measured branch lengths using d_S values. We separated each gene family and compared the rates of evolution of individual genes across all ten *Drosophila* species. There does not appear to be any directionality to observed variations in divergence rates, though it is noteworthy that *dan* and *danr* appear to have higher rates of evolution in all species (based on branch length). Therefore, individual genes are just as likely to diverge more rapidly within the *melanogaster* and *obscura* subgroups as they are to diverge more slowly. We therefore conclude that there are no species or species subgroups that have particularly high rates of evolution for the network as a whole (Fig. 2).

Paralogs in the network are evolving at different rates

We then qualitatively estimated when each of the five sets of duplicate gene pairs that exist within the RD network may have arisen, and also compared rates of evolution between the paralogs. The Pax6 homologs, *ey* and *toy*, are present within all ten *Drosophilids* and the three basal insect genomes (Fig. 3A, data not shown). We observe longer branch lengths for genes within the EY clade indicating a faster rate of sequence evolution for *ey* genes as compared to *toy* (Wilcoxon test; p=0.00016). These differing rates are in line with experimental evidence that Ey and Toy proteins have different functions within the eye and are evolving different transcriptional activities (Czerny et al., 1999;Punzo et al., 2004;Quiring et al., 1994;Weasner et al., 2009). These differences include a stronger transcriptional activation domain and a repressor domain for Ey (Weasner et al., 2009).

The two Pax6(5a) genes *eyg* and *toe* are present in all ten *Drosophila* species but only a single gene is found in *Aedes aegypti*, *Tribolium castaneum* and *Apis mellifera*. Thus we infer that the duplication of ancestral Pax6(5a) to yield *eyg* and *toe* occurred sometime prior to the diversification of the *Drosophilid* lineage (Fig. 3B, data not shown, Bao and Friedrich, 2009). Relative to Pax6, this duplication appears to have been more recent. In contrast to several other duplicate genes within the network, the rates of divergence for *eyg* and *toe* are not significantly different from each other (Wilcoxon test; p=0.5). The similar divergence rates for *eyg* and *toe* (at the whole gene level) appear to be supported by the fact that Eyg and Toe proteins are thought to play somewhat redundant roles in the eye (Yao et al., 2008).

Our gene tree analysis indicates that the duplication of the ancestral Tsh/Tio gene also occurred before the diversification of the *Drosophilids*. This is based on the clear identification of a single *tsh/tio* gene in the basal insects while finding both *tsh* and *tio* in all ten *Drosophila* species (Fig. 3C, data not shown, Bao and Friedrich, 2009;Shippy et al., 2008). We have observed that the branch lengths of members of the TSH clade are longer than those of the TIO clade indicating a faster rate of sequence evolution for *tsh* class genes (Wilcoxon rank sum test with continuity correction; W=0, p=0.00017). It should be noted that the ancestral Tsh/Tio protein contains four zinc finger domains. *Drosophila* Tio shares this structure while the Tsh proteins have only three such motifs. In the developing eye both Tsh and Tio proteins are distributed in similar patterns, at nearly identical levels and appear to be at least partially redundant (Bessa et al., 2009;Datta et al., 2009;Laugier et al., 2005). However, there are significant differences in the way that the two genes induce ectopic eye formation and promote cell proliferation (Datta et al., 2009).

The duplication events that gave rise to *so, optix* and *DSix4* predate the diversification of the species that we have used in this study and thus represent the most ancient set of duplications within the known RD network (Fig. 3E). Sequence comparisons indicate that the ancestral SIX gene likely duplicated to produce *so* and a *DSix4/optix* intermediate which subsequently duplicated to give rise to the modern day *DSix4* and *optix* genes. Perhaps due to the extreme diversity in SIX protein function, all three pair wise comparisons of standardized branch lengths are significantly different, suggesting that all members of the SIX family are diverging at different rates (Wilcoxon test; $p<0.05$). In regards to retinal development, the So and Optix proteins have distinct effects on transcription with So functioning primarily as an activator via binding to Eya while Optix serves as a repressor through interactions with Groucho (Gro: Kenyon et al., 2005a;Kenyon et al., 2005b;Pignoni et al., 1997). It also appears that sequences within the C-terminal segments of the SIX proteins further distinguish So and Optix (Weasner et al., 2009).

The paralogs *dan* and *danr* also have duplicated prior to *Drosophila* diversification. A single copy is present in basal insects while all ten *Drosophila* species have both genes (Fig. 3D, data not shown, Bao and Friedrich, 2009). Similar to *eyg* and *toe*, branch lengths are not significantly longer for *dan* than *danr* within *Drosophila* (Wilcoxon test, $p=0.684$). A structure/function analysis for these two paralogs has yet to be performed. However, our region-by-region analysis of d_N/d_S appears to provide some clues as to where some local differences may exist (see below).

Selection signatures and purifying selection on retinal genes

To determine the relative selective constraint on each of the RD network genes and to control for mutation rate, we calculated the ratio of non-synonymous to synonymous substitutions (d_N/d_S) using sequences from species in the *melanogaster* subgroup. We find that all members of this network are under varying degrees of purifying selection. For example, *toe, danr* and *ey* have the highest d_N/d_S ratios (0.31, 0.24 and 0.22) while *hth* and *dSix4* have the lowest rates of substitution (0.031 and 0.033; Figure 5A, Table 1). We also observe that duplicate genes have disparate and statistically significant d_N/d_S ratios (95% confidence intervals for each duplicate gene were obtained by 1000 bootstrap replicates, Fig. 5A, Table 1). The d_N/d_S values for the Pax genes *ey* and *toy* are 0.22 (.19, .27) and 0.06 {. 06, .09} while for *eyg* and *toe* they are 0.11 (.05, .21) and 0.31 (.23, .32). The values for the Tsh/Tio genes *tsh* and *tio* are 0.07 (.06, .09) and 0.17 (.16, .18) respectively while those for the SIX genes *so* and *optix* are 0.14 (.11, .16) and 0.08 (.07, .11). In contrast, this trend does not hold true for *dan* and *danr*, which, while having ratios of 0.18 and 0.244 also have overlapping confidence intervals (*dan*: .14954, .21265; *danr*: 08112, .32509). However, the bootstrap values themselves are suggestive of qualitative differences and *danr* has high variance around the mean.

We were interested in determining if any RD genes were under positive selection. Upon using population sequence data from 39 published *D. melanogaster* lines we do not find any such evidence (MK tests, $p>0.05$). Of all the genes only *nemo* had an excess of non-synonymous fixed differences between species, which would be consistent with positive selection. However, it was only weakly significant using a Fisher's exact test ($p=0.066$). For all genes, there was a scarcity of polymorphisms and this small number of differences across at least one row or column in the 2×2 contingency tables of polymorphism and divergence made it difficult to assess significance. Across the 39 lines we never find more than 4 non-synonymous polymorphisms for any gene in our sample and the average number of total polymorphisms (synonymous + non-synonymous) is just over 5. Therefore, our population analysis is consistent with interspecific analysis of d_N/d_S, whereby the genes in this network experience varying degrees of purifying selection as opposed to recurrent positive selection.

Differential Selection across Protein-Coding Regions

Our data suggest that each of the paralogs is under varying degrees of purifying selection (Fig. 5A, Table 1). We next set out to determine the constraint profiles for each gene by calculating the d_N/d_S ratios for sections of each gene that code for either known functional domains or for non-conserved portions of the protein that hitherto have no ascribed activity. For both the full-length genes and individual domains, there is little to no correlation between length (measured in number of base pairs) and d_N/d_S ratios (full-length: $y=0x+0.1$, $R^2=.01$, domains: $y=0x+0.3$, $R^2=.15$). However, as length of region decreases, the standard error obtained through bootstrapping increases greatly, as expected. We gain some power to analyze particular domains by pooling across genes. Overall, we find high variation among domains both within and between genes but also find some very predictable patterns.

Of the fourteen genes that constitute the known RD network, twelve encode DNA binding proteins. We examined the d_N/d_S ratios for the four different types of DNA binding domains (paired, homeobox, zinc finger and pipsqueak) that are found within these proteins as well as *DSix4* (Fig. 4, Table 1). Our analysis of these domains indicates that they are under strong purifying selection with several domains completely conserved between species. In fact, the highest d_N/d_S values recorded for any DNA binding domain within the network is just 0.1 (Hth homeodomain [not shown] and Tio zinc finger #4; Table 1). By pooling domains across genes, however, we find that DNA binding domains have significantly lower values of d_N/d_S than non-DNA binding domains (Student's t-test; $t=4.37$, $p=0.0001$). Further, when we include protein-protein interaction domains such as the SIX domains of *so*, *optix* and *DSix4*, we find significantly lower d_N/d_S values compared with all non-conserved domains ($t=4.64$, $p<.00001$). These conserved domains are exactly the sorts of regions we expect to have low rates of divergence across taxa. The non-conserved segments are expected to have higher d_N/d_S ratios (Fig. 4) and are likely to be the areas in which ancestral functions are being lost or new activities are being gained (see below).

Differential Selection: A Comparison of Paralogs

A comparison of domains within paralog pairs offers the opportunity to identify areas of potential sub-functionalization and neo-functionalization. Overall, the d_N/d_S ratios for the non-conserved segments of each protein indicated that these regions are under variable degrees of relaxed selection relative to conserved segments (Fig. 4) with values, in some cases, approaching 0.9 (Toe CT region; Table 1). Significantly, we are able to correlate the regions within the largest variations with recently identified functional differences between each paralog pair. In the majority of cases the largest variations are seen in the non-conserved portions of the proteins while the DNA binding and protein-protein interaction domains appear to under the strongest purifying selection (Fig. 4, 5B-L, Table 1). We first examined the Pax6 genes *ey* and *toy*. It has been noted that Ey appears to be able to promote ectopic eye formation in a broader range of tissues than Toy (Czerny et al., 1999;Halder et al., 1995; Salzer and Kumar, 2010). One of the areas with the largest difference in d_N/d_S maps to the C-terminal (CT) region (Fig. 5B,C;Supplementary Fig.1). A recent structure/function analysis of these Pax6 proteins indicates that the CT segment of Ey has a transcriptional activation domain that is significantly stronger than the one found within the CT of Toy (Weasner et al., 2009). That same study also identified a putative repressor domain within the region of Ey that links the two DNA binding domain (B). This activity appears to be absent from the Toy protein (Weasner et al., 2009). Our analysis here indicates that the largest differences in d_N/d_S maps to this linker region (.087 vs. .27; Fig. 5B,C, Table 1).

We then analyzed the levels of selection across the Pax6(5a) genes *eyg* and *toe* and find that the highest variation in d_N/d_S ratios maps to the B and CT regions (Fig. 5D,E, Table 1). Eyg

and Toe are expressed in nearly identical patterns in the developing eye, are functionally redundant and both serve as transcriptional repressors (Yao and Sun, 2005; Yao et al., 2008). However, mechanistic differences between how the two proteins influence transcription were experimentally identified by molecular dissections of the paralog proteins (Yao and Sun, 2005; Yao et al., 2008). These studies identified two repressor domains residing within the B and CT portions of Eyg (Yao and Sun, 2005) but only a single repressor domain within the B of Toe (Yao et al., 2008). In addition to the differences in d_N/d_S ratios of these two segments we also note with interest that the absolute d_N/d_S value of the *toe* CT is measured at 0.89, which is approaching the 1.0 threshold for positive selection (Fig. 5D,E, Table 1, Suppl. Fig. 1). An analysis using 100 base pair windows on the Ey, Toy, Eyg and Toe CTs reveals areas with higher and lower d_N/d_S ratios around the reported mean (Suppl. Fig. 1). This smaller window size reveals more localized changes within the larger non-conserved domain, which can be further analyzed functionally.

Early eye formation is also dependent upon members of the SIX family of homeobox transcription factors. Of the three genes that are present in flies only the So and Optix proteins function during retinal development. In addition to high sequence conservation within the DNA binding and protein-protein interaction domains, recent reports have indicated that both proteins bind to nearly identical DNA sequences (Berger et al., 2008; Noyes et al., 2008) and can bind to a common set of protein co-factors (Kenyon et al., 2005a; Kenyon et al., 2005b). However, rescue experiments indicate that these genes are not functionally interchangeable (Weasner et al., 2007). Our analysis of selection pressures across the SIX genes indicates that the region under the most relaxed constraints is the CT segment of *so*, which has a d_N/d_S value greater than that of either *optix* or *DSix4* (Fig. 5H-J, Table 1). The CT segments were recently shown to contribute to the functional differences between the So and Optix proteins (Weasner and Kumar, 2009).

The Tsh and Tio protein paralogs are structurally different than any of the other RD proteins in the fact that they both contain differing numbers of DNA binding domains. Tsh contains three zinc finger domains while Tio has three such motifs. While this structural difference could account for some reported functional differences (Datta et al., 2009) we set out to determine if other regions of these paralogs could also be acquiring new or losing old functions. Our analysis of selection pressures indicates that there are significant differences in the d_N/d_S values for the N-terminal (NT) segment as well as the first and third zinc finger domains (Fig. 5F,G, Table 1). This is particularly interesting as it represents the only paralog pair in which the conserved DNA binding domains have significant differences in the d_N/d_S values. Functional dissections of these proteins indicate that some differences in the abilities of these proteins to induce cell proliferation and support eye development reside within these domains (R.R. Datta and J.P. Kumar, unpublished data).

Finally, an analysis of the last paralog pair, *dan* and *danr*, indicates that the largest disparity in d_N/d_S values is within the NT segment (Fig. 5K,L, Table 1). Unfortunately, structure/function data do not yet exist for this set of duplicate genes. Based on our analysis of the other four duplicate gene pairs we predict that any functional differences that exist between the Dan and Danr proteins will be attributable to either the loss of old function or the acquisition of new ones within the NT segment.

Identification of Important Residues in Structured Domains

We have identified several positions within the RD network proteins that, throughout the *Drosophilid* lineage, are occupied by one amino acid in one paralog but by another residue in the other paralog. Such positions are considered “clade specific” residues (Table 2, Suppl. Fig. 2). The underlying substitution events within the genome that give rise to these features are predicted to have occurred prior to the diversification of the *Drosophilids*. Position 7 of

the Pax6 PD exemplifies a clade specific residue: it is occupied by either a valine (Ey) or isoleucine (Toy) in all species. We have also identified several residues that we consider "group specific" amino acids (Table 2, Suppl. Fig. 2). These marked substitution events that have occurred relatively recently and can be only found within a small subset of species. For instance, threonine and glutamic acid residues occupy positions 8 and 10 respectively in nearly all Ey and Toy HDs. However, species within the *melanogaster* subgroup have a serine at position 8 (Ey) and an aspartic acid at residue 10 (Toy). These small scale changes on the critical domains of the network may account for the varying degrees of selection inferred from the constrained domains and in part, for the changes in DNA binding and protein interacting activities of the duplicate gene pairs. Interestingly, Dan and Danr stand out in that they show minimal clade or group-specific changes in the PSQ DNA binding domain.

Discussion

The evolutionary conserved retinal determination (RD) network governs early decisions in eye development in a broad spectrum of organisms that range from insects such as *Drosophila* to mammals such as humans. Maintaining the functional integrity of such multipurpose networks is critical. Genes that are pleiotropic are expected to be under stringent purifying selection as there is the additional pressure of numerous cellular and developmental processes being regulated. In flies, where this network was first identified, it controls the development of learning and memory centers of the brain, several mesodermal derivatives, the gonads and select cells within the central nervous system (Bai and Montell, 2002; Bonini et al., 1998; Callaerts et al., 2001; Chang et al., 2003; Fabrizio et al., 2003; Kammermeier et al., 2001; Kurusu et al., 2000; Mardon et al., 1994; Niimi et al., 2002; Noveen et al., 2000). In vertebrates, the RD network regulates ear, nose, kidney, and muscle specification in vertebrates (Brodbeck and Englert, 2004; Gong et al., 2007; Hammond et al., 1998; Hanson, 2001; Heanue et al., 1999; Kalatzis et al., 1998; Laclef et al., 2003; Relaix and Buckingham, 1999; Simpson and Price, 2002; Xu et al., 2003). Together, the wide range of developmental effects and disease states make the RD network arguably one of the best-studied gene regulatory networks in development.

In this paper we performed an evolutionary analysis on each member of the network within ten *Drosophila* species as well as in *A. gambiae*, *T. castaneum* and *A. mellifera*. In particular, we focused on identifying when duplication events within the network took place, the rate at which each paralog evolved in relation to one another and the selection signatures across the functional conserved domains and non-conserved segments. We observe that the network as a whole is constrained across all ten *Drosophila* species. We do find that amongst each pair of duplicate genes, the paralogs are evolving at different rates suggesting that they may be undergoing either sub or neo-functionalization. We extended these findings by calculating d_N/d_S values across the coding regions for each paralog pair. Predictably, the d_N/d_S ratios for the functionally conserved domains are significantly lower than that of the non-conserved segments, confirming our hypothesis that there is differential selection acting on genes in the RD network. However, we also found that the d_N/d_S values for the non-conserved regions could vary significantly. Upon closer inspection, the non-conserved segments with the greatest differences in d_N/d_S ratios appear to be the regions of the gene that have been shown experimentally to be have gained or lost functions. If our d_N/d_S analysis preceded published structure/function studies we would have accurately predicted the location of functional differences within the Ey/Toy, Eyg/Toe, Six/Optix and Tsh/Tio gene pairs. Based on these correlations we suggest that our methodology can be used to accurately identify evolving regions of proteins, particularly those that are encoded by duplicate genes.

The potential usefulness of such a method is a valid consideration. Structure/function analyses are time-consuming and laborious, and if performed in model systems such as the mouse can also be prohibitively expensive. The approach presented here represents new way to look at duplicate genes and make very accurate predictions as to which deletion and chimeric constructs would be the most informative in terms of identifying new functional domains and activities. Of the four duplicate gene pairs within the retinal determination network that have been subjected to molecular dissections (Weasner et al., 2007; Weasner et al., 2009; Yao et al., 2008) each of the regions that were experimentally identified as evolving new functions are accurately predicted by the approach described here. We therefore predict that this will be a useful tool for analyzing other duplicate gene pairs as well as gene families.

The number of gene pairs or gene families that could be subjected to this analysis is extensive. For example, just within the developing Drosophila eye there are the *spalt major* (*salm*) and *spalt related* (*salr*) genes that govern the specification of the R3/4 photoreceptors (Domingos et al., 2004), the *BarH1* and *BarH2* homeobox genes that regulate development of the R1/6 pair of photoreceptors (Hayashi et al., 1998), the six genes that code for light capturing rhodopsin proteins (Morante et al., 2007) and a pair of paralogs that code for the Trp and Trpl channels (Harteneck et al., 2000). Additionally, there are several hundred genes that constitute the olfactory and gustatory gene families in flies (Keller and Vosshall, 2003; Scott et al., 2001). These represent just a few examples of the types of duplicate genes and gene families that could be subjected to our analysis prior to the initiation of molecular dissections of protein function. Our methodology could have a greater impact on studies involving duplicate genes in vertebrates such as the mouse. The whole genome duplication event that occurred during early vertebrate development has often complicated gene evolution studies. For example, the members of the vertebrate RD network are present in multiple copies compared to the *Drosophila* genes (Hanson, 2001; Kumar, 2009a; Wawersik and Maas, 2000). Due to the laborious and expensive nature of doing *in vivo* structure/function assays in vertebrate systems, particularly the mouse, very few molecular dissections of vertebrate genes are conducted *in vivo*. We propose that the method described here could be a valuable resource in pinpointing the regions of duplicate genes that are most likely evolving new functions, thereby distinguishing one paralog from its sister gene. It is also likely that this kind of differential selection allows the functional integrity of the gene to be maintained so as not to compromise the regulatory networks while leaving room for new interactions and possible sub and neo-functionalization. Studying paralogs in this way will likely also shed considerable light on network evolution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Mike Wade, Phil Nista and Bonnie Weasner for comments on the manuscript and Brandon Weasner for help on Figure 1A. This work was supported by a grant from the National Eye Institute (R01 EY014863) to Justin P. Kumar.

References

Aldaz S, Morata G, Azpiazu N. The Pax-homeobox gene eyegone is involved in the subdivision of the thorax of Drosophila. Development. 2003; 130:4473–82. [PubMed: 12900462]

Amoutzias GD, Robertson DL, Oliver SG, Bornberg-Bauer E. Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. EMBO Rep. 2004; 5:274–9. [PubMed: 14968135]

Bai J, Montell D. Eyes absent, a key repressor of polar cell fate during Drosophila oogenesis. Development. 2002; 129:5377–88. [PubMed: 12403709]

Balakirev ES, Ayala FJ. Pseudogenes: are they “junk” or functional DNA? Annu Rev Genet. 2003; 37:123–51. [PubMed: 14616058]

Bao R, Friedrich M. Molecular evolution of the Drosophila retinome: exceptional gene gain in the higher Diptera. Mol Biol Evol. 2009

Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET, Khalid F, Zhang W, Newburger D, Jaeger SA, Morris QD, Bulyk ML, Hughes TR. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell. 2008; 133:1266–76. [PubMed: 18585359]

Bessa J, Carmona L, Casares F. Zinc-finger paralogues tsh and tio are functionally equivalent during imaginal development in Drosophila and maintain their expression levels through auto- and cross-negative feedback loops. Dev Dyn. 2009; 238:19–28. [PubMed: 19097089]

Bonini NM, Bui QL, Gray-Board GL, Warrick JM. The Drosophila eyes absent gene directs ectopic eye formation in a pathway conserved between flies and vertebrates. Development. 1997; 124:4819–4826. [PubMed: 9428418]

Bonini NM, Leiserson WM, Benzer S. The eyes absent gene: genetic control of cell survival and differentiation in the developing Drosophila eye. Cell. 1993; 72:379–95. [PubMed: 8431945]

Bonini NM, Leiserson WM, Benzer S. Multiple roles of the eyes absent gene in Drosophila. Dev Biol. 1998; 196:42–57. [PubMed: 9527880]

Braid LR, Verheyen EM. Drosophila nemo promotes eye specification directed by the retinal determination gene network. Genetics. 2008; 180:283–99. [PubMed: 18757943]

Brodbeck S, Englert C. Genetic determination of nephrogenesis: the Pax/Eya/Six gene network. Pediatr Nephrol. 2004; 19:249–55. [PubMed: 14673635]

Callaerts P, Halder G, Gehring WJ. PAX-6 in development and evolution. Annu Rev Neurosci. 1997; 20:483–532. [PubMed: 9056723]

Callaerts P, Leng S, Clements J, Benassayag C, Cribbs D, Kang YY, Walldorf U, Fischbach KF, Strauss R. Drosophila Pax-6/eyeless is essential for normal adult brain structure and function. J Neurobiol. 2001; 46:73–88. [PubMed: 11153010]

Chang T, Younossi-Hartenstein A, Hartenstein V. Development of neural lineages derived from the sine oculis positive eye field of Drosophila. Arthropod Struct Dev. 2003; 32:303–17. [PubMed: 18089014]

Chen CC, Li WH, Sung HM. Patterns of internal gene duplication in the course of metazoan evolution. Gene. 2007; 396:59–65. [PubMed: 17442504]

Cheyette BN, Green PJ, Martin K, Garren H, Hartenstein V, Zipursky SL. The Drosophila sine oculis locus encodes a homeodomain-containing protein required for the development of the entire visual system. Neuron. 1994; 12:977–96. [PubMed: 7910468]

Choi KW, Benzer S. Rotation of photoreceptor clusters in the developing Drosophila eye requires the nemo gene. Cell. 1994; 78:125–36. [PubMed: 8033204]

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, Pollard DA, Sackton TB, Larracuente AM, Singh ND, Abad JP, Abt DN, Adryan B, Aguade M, Akashi H, Anderson WW, Aquadro CF, Ardell DH, Arguello R, Artieri CG, Barbash DA, Barker D, Barsanti P, Batterham P, Batzoglou S, Begun D, Bhutkar A, Blanco E, Bosak SA, Bradley RK, Brand AD, Brent MR, Brooks AN, Brown RH, Butlin RK, Caggese C, Calvi BR, Bernardo de Carvalho A, Caspi A, Castrezana S, Celniker SE, Chang JL, Chapple C, Chatterji S, Chinwalla A, Civetta A, Clifton SW, Comeron JM, Costello JC, Coyne JA, Daub J, David RG, Delcher AL, Delehaunty K, Do CB, Ebling H, Edwards K, Eickbush T, Evans JD, Filipski A, Findeiss S, Freyhult E, Fulton L, Fulton R, Garcia AC, Gardiner A, Garfield DA, Garvin BE, Gibson G, Gilbert D, Gnerre S, Godfrey J, Good R, Gotea V, Gravely B, Greenberg AJ, Griffiths-Jones S, Gross S, Guigo R, Gustafson EA, Haerty W, Hahn MW, Halligan DL, Halpern AL, Halter GM, Han MV, Heger A, Hillier L, Hinrichs AS, Holmes I, Hoskins RA, Hubisz MJ, Hultmark D, Huntley MA, Jaffe DB, Jagadeeshan S, Jeck WR, Johnson J, Jones CD, Jordan WC, Karpen GH, Kataoka E, Keightley PD, Kheradpour P, Kirkness EF, Koerich LB, Kristiansen K, Kudrna D, Kulathinal RJ, Kumar S, Kwok R, Lander E, Langley CH, Lapoint R,

Lazzaro BP, Lee SJ, Levesque L, Li R, Lin CF, Lin MF, Lindblad-Toh K, Llopart A, Long M, Low L, Lozovsky E, Lu J, Luo M, Machado CA, Makalowski W, Marzo M, Matsuda M, Matzkin L, McAllister B, McBride CS, McKernan B, McKernan K, Mendez-Lago M, Minx P, Mollenhauer MU, Montooth K, Mount SM, Mu X, Myers E, Negre B, Newfeld S, Nielsen R, Noor MA, O'Grady P, Pachter L, Papaceit M, Parisi MJ, Parisi M, Parts L, Pedersen JS, Pesole G, Phillippy AM, Ponting CP, Pop M, Porcelli D, Powell JR, Prohaska S, Pruitt K, Puig M, Quesneville H, Ram KR, Rand D, Rasmussen MD, Reed LK, Reenan R, Reily A, Remington KA, Rieger TT, Ritchie MG, Robin C, Rogers YH, Rohde C, Rozas J, Rubenfield MJ, Ruiz A, Russo S, Salzberg SL, Sanchez-Gracia A, Saranga DJ, Sato H, Schaeffer SW, Schatz MC, Schlenke T, Schwartz R, Segarra C, Singh RS, Sirot L, Sirota M, Sisneros NB, Smith CD, Smith TF, Spieth J, Stage DE, Stark A, Stephan W, Strausberg RL, Strempel S, Sturgill D, Sutton G, Sutton GG, Tao W, Teichmann S, Tobari YN, Tomimura Y, Tsolas JM, Valente VL, Venter E, Venter JC, Vicario S, Vieira FG, Vilella AJ, Villasante A, Walenz B, Wang J, Wasserman M, Watts T, Wilson D, Wilson RK, Wing RA, Wolfner MF, Wong A, Wong GK, Wu CI, Wu G, Yamamoto D, Yang HP, Yang SP, Yorke JA, Yoshida K, Zdobnov E, Zhang P, Zhang Y, Zimin AV, Baldwin J, Abdouelleil A, Abdulkadir J, Abebe A, Abera B, Abreu J, Acer SC, Aftuck L, Alexander A, An P, Anderson E, Anderson S, Arachi H, Azer M, Bachantsang P, Barry A, Bayul T, Berlin A, Bessette D, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukhgalter B, Bourzgui I, Brown A, Cahill P, Channer S, Cheshatsang Y, Chuda L, Citroen M, Collymore A, Cooke P, Costello M, D'Aco K, Daza R, De Haan G, DeGray S, DeMaso C, Dhargay N, Dooley K, Dooley E, Doricent M, Dorje P, Dorjee K, Dupes A, Elong R, Falk J, Farina A, Faro S, Ferguson D, Fisher S, Foley CD, Franke A, Friedrich D, Gadbois L, Gearin G, Gearin CR, Giannoukos G, Goode T, Graham J, Grandbois E, Grewal S, Gyaltsen K, Hafez N, Hagos B, Hall J, Henson C, Hollinger A, Honan T, Huard MD, Hughes L, Hurhula B, Husby ME, Kamat A, Kanga B, Kashin S, Khazanovich D, Kisner P, Lance K, Lara M, Lee W, Lennon N, Letendre F, LeVine R, Lipovsky A, Liu X, Liu J, Liu S, Lokyitsang T, Lokyitsang Y, Lubonja R, Lui A, MacDonald P, Magnisalis V, Maru K, Matthews C, McCusker W, McDonough S, Mehta T, Meldrim J, Meneus L, Mihai O, Mihalev A, Mihova T, Mittelman R, Mlenga V, Montmayeur A, Mulrain L, Navidi A, Naylor J, Negash T, Nguyen T, Nguyen N, Nicol R, Norbu C, Norbu N, Novod N, O'Neill B, Osman S, Markiewicz E, Oyono OL, Patti C, Phunkhang P, Pierre F, Priest M, Raghuraman S, Rege F, Reyes R, Rise C, Rogov P, Ross K, Ryan E, Settipalli S, Shea T, Sherpa N, Shi L, Shih D, Sparrow T, Spaulding J, Stalker J, Stange-Thomann N, Stavropoulos S, Stone C, Strader C, Tesfaye S, Thomson T, Thoulutsang Y, Thoulutsang D, Topham K, Topping I, Tsamla T, Vassiliev H, Vo A, Wangchuk T, Wangdi T, Weiand M, Wilkinson J, Wilson A, Yadav S, Young G, Yu Q, Zembek L, Zhong D, Zimmer A, Zwirko Z, Alvarez P, Brockman W, Butler J, Chin C, Grabherr M, Kleber M, Mauceli E, MacCallum I. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007; 450:203–18. [PubMed: 17994087]

Curtiss J, Burnett M, Mlodzik M. distal antenna and distal antenna-related function in the retinal determination network during eye development in Drosophila. Dev Biol. 2007; 306:685–702. [PubMed: 17493605]

Czerny T, Halder G, Kloter U, Souabni A, Gehring WJ, Busslinger M. twin of eyeless, a second Pax-6 gene of Drosophila, acts upstream of eyeless in the control of eye development. Mol Cell. 1999; 3:297–307. [PubMed: 10198632]

Datta RR, Lurye JM, Kumar JP. Restriction of ectopic eye formation by Drosophila teashirt and tiptop to the developing antenna. Dev Dyn. 2009

Dermitzakis ET, Clark AG. Differential selection after duplication in mammalian developmental genes. Mol Biol Evol. 2001; 18:557–62. [PubMed: 11264407]

Domingos PM, Mlodzik M, Mendes CS, Brown S, Steller H, Mollereau B. Spalt transcription factors are required for R3/R4 specification and establishment of planar cell polarity in the Drosophila eye. Development. 2004; 131:5695–702. [PubMed: 15509769]

Donner AL, Maas RL. Conservation and non-conservation of genetic pathways in eye specification. Int J Dev Biol. 2004; 48:743–53. [PubMed: 15558467]

Fabrizio JJ, Boyle M, DiNardo S. A somatic role for eyes absent (eya) and sine oculis (so) in Drosophila spermatocyte development. Dev Biol. 2003; 258:117–28. [PubMed: 12781687]

Fasano L, Roder L, Core N, Alexandre E, Vola C, Jacq B, Kerridge S. The gene teashirt is required for the development of Drosophila embryonic trunk segments and encodes a protein with widely spaced zinc finger motifs. Cell. 1991; 64:63–79. [PubMed: 1846092]

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999; 151:1531–45. [PubMed: 10101175]

Gardiner A, Barker D, Butlin RK, Jordan WC, Ritchie MG. Drosophila chemoreceptor gene evolution: selection, specialization and genome size. Mol Ecol. 2008; 17:1648–57. [PubMed: 18371013]

Gehring WJ. The master control gene for morphogenesis and evolution of the eye. Genes Cells. 1996; 1:11–5. [PubMed: 9078363]

Gehring WJ, Ikeo K. Pax 6: mastering eye morphogenesis and eye evolution. Trends Genet. 1999; 15:371–7. [PubMed: 10461206]

Gertz EM, Yu YK, Agarwala R, Schaffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 2006; 4:41. [PubMed: 17156431]

Gibert JM. The evolution of engrailed genes after duplication and speciation events. Dev Genes Evol. 2002; 212:307–18. [PubMed: 12185484]

Gong KQ, Yallowitz AR, Sun H, Dressler GR, Wellik DM. A Hox-Eya-Pax complex regulates early kidney developmental gene expression. Mol Cell Biol. 2007; 27:7661–8. [PubMed: 17785448]

Gu Z, Rifkin SA, White KP, Li WH. Duplicate genes increase gene expression diversity within and between species. Nat Genet. 2004; 36:577–9. [PubMed: 15122255]

Halder G, Callaerts P, Gehring WJ. Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. Science. 1995; 267:1788–92. [PubMed: 7892602]

Hammond KL, Hanson IM, Brown AG, Lettice LA, Hill RE. Mammalian and Drosophila dachshund genes are related to the Ski protooncogene and are expressed in eye and limb. Mech Dev. 1998; 74:121–31. [PubMed: 9651501]

Hanson IM. Mammalian homologues of the Drosophila eye specification genes. Semin Cell Dev Biol. 2001; 12:475–84. [PubMed: 11735383]

Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M. Identification of pseudogenes in the Drosophila melanogaster genome. Nucleic Acids Res. 2003; 31:1033–7. [PubMed: 12560500]

Harteneck C, Plant TD, Schultz G. From worm to man: three subfamilies of TRP channels. Trends Neurosci. 2000; 23:159–66. [PubMed: 10717675]

Hayashi T, Kojima T, Saigo K. Specification of primary pigment cell and outer photoreceptor fates by BarH1 homeobox gene in the developing Drosophila eye. Dev Biol. 1998; 200:131–45. [PubMed: 9705222]

Heanue TA, Reshef R, Davis RJ, Mardon G, Oliver G, Tomarev S, Lassar AB, Tabin CJ. Synergistic regulation of vertebrate muscle development by Dach2, Eya2, and Six1, homologs of genes required for Drosophila eye formation. Genes Dev. 1999; 13:3231–43. [PubMed: 10617572]

Hughes AL. The evolution of functionally novel proteins after gene duplication. Proc Biol Sci. 1994; 256:119–24. [PubMed: 8029240]

Hughes AL, Friedman R. Gene duplication and the properties of biological networks. J Mol Evol. 2005; 61:758–64. [PubMed: 16315107]

Hurley I, Hale ME, Prince VE. Duplication events and the evolution of segmental identity. Evol Dev. 2005; 7:556–67. [PubMed: 16336409]

Jean D, Ewan K, Gruss P. Molecular regulators involved in vertebrate eye development. Mech. Dev. 1998; 76:3–18. [PubMed: 9767078]

Jun S, Wallen RV, Goriely A, Kalionis B, Desplan C. Lune/eye gone, a Pax-like protein, uses a partial paired domain and a homeodomain for DNA recognition. Proc Natl Acad Sci U S A. 1998; 95:13720–5. [PubMed: 9811867]

Kalatzis V, Sahly I, El-Amraoui A, Petit C. Eya1 expression in the developing ear and kidney: towards the understanding of the pathogenesis of Branchio-Oto-Renal (BOR) syndrome. Dev Dyn. 1998; 213:486–99. [PubMed: 9853969]

Kammermeier L, Leemans R, Hirth F, Flister S, Wenger U, Walldorf U, Gehring WJ, Reichert H. Differential expression and function of the Drosophila Pax6 genes eyeless and twin of eyeless in

embryonic central nervous system development. Mech Dev. 2001; 103:71–8. [PubMed: 11335113]

Keller A, Vosshall LB. Decoding olfaction in Drosophila. Curr Opin Neurobiol. 2003; 13:103–10. [PubMed: 12593988]

Kenyon KL, Li DJ, Clouser C, Tran S, Pignoni F. Fly SIX-type homeodomain proteins Sine oculis and Optix partner with different cofactors during eye development. Dev Dyn. 2005a; 234:497–504. [PubMed: 15937930]

Kenyon KL, Yang-Zhou D, Cai CQ, Tran S, Clouser C, Decene G, Ranade S, Pignoni F. Partner specificity is essential for proper function of the SIX-type homeodomain proteins Sine oculis and Optix during fly eye development. Dev Biol. 2005b; 286:158–68. [PubMed: 16125693]

Kozmik Z, Holland ND, Kreslova J, Oliveri D, Schubert M, Jonasova K, Holland LZ, Pestarino M, Benes V, Candiani S. Pax-Six-Eya-Dach network during amphioxus development: conservation in vitro but context specificity in vivo. Dev Biol. 2007; 306:143–59. [PubMed: 17477914]

Krakauer DC, Nowak MA. Evolutionary preservation of redundant duplicated genes. Semin Cell Dev Biol. 1999; 10:555–9. [PubMed: 10597640]

Kumar JP. Signalling pathways in Drosophila and vertebrate retinal development. Nat Rev Genet. 2001; 2:846–57. [PubMed: 11715040]

Kumar JP. The molecular circuitry governing retinal determination. Biochim Biophys Acta. 2009a; 1789:306–14. [PubMed: 19013263]

Kumar JP. The sine oculis homeobox (SIX) family of transcription factors as regulators of development and disease. Cell Mol Life Sci. 2009b; 66:565–83. [PubMed: 18989625]

Kurusu M, Nagao T, Walldorf U, Flister S, Gehring WJ, Furukubo-Tokunaga K. Genetic control of development of the mushroom bodies, the associative learning centers in the Drosophila brain, by the eyeless, twin of eyeless, and Dachshund genes. Proc Natl Acad Sci U S A. 2000; 97:2140–4. [PubMed: 10681433]

Laclef C, Souil E, Demignon J, Maire P. Thymus, kidney and craniofacial abnormalities in Six 1 deficient mice. Mech Dev. 2003; 120:669–79. [PubMed: 12834866]

Laugier E, Yang Z, Fasano L, Kerridge S, Vola C. A critical role of teashirt for patterning the ventral epidermis is masked by ectopic expression of tiptop, a paralog of teashirt in Drosophila. Dev Biol. 2005; 283:446–58. [PubMed: 15936749]

Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000; 290:1151–5. [PubMed: 11073452]

Lynch M, Conery JS. The evolutionary demography of duplicate genes. J Struct Funct Genomics. 2003; 3:35–44. [PubMed: 12836683]

Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. Genetics. 2000; 154:459–73. [PubMed: 10629003]

Mardon G, Solomon NM, Rubin GM. dachshund encodes a nuclear protein required for normal eye and leg development in Drosophila. Development. 1994; 120:3473–86. [PubMed: 7821215]

Morante J, Desplan C, Celik A. Generating patterned arrays of photoreceptors. Curr Opin Genet Dev. 2007; 17:314–9. [PubMed: 17616388]

Nei, M.; Kumar, S. Molecular Evolution and Phylogenetics. Oxford University Press; US: 2000.

Niimi T, Clements J, Gehring WJ, Callaerts P. Dominant-negative form of the Pax6 homolog eyeless for tissue-specific loss-of-function studies in the developing eye and brain in drosophila. Genesis. 2002; 34:74–5. [PubMed: 12324952]

Noveen A, Daniel A, Hartenstein V. Early development of the Drosophila mushroom body: the roles of eyeless and dachshund. Development. 2000; 127:3475–88. [PubMed: 10903173]

Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell. 2008; 133:1277–89. [PubMed: 18585360]

Ohno, S. Evolution by Gene Duplication. Springer-Verlag; Berlin-Heidelberg-New York: 1970.

Ohta T. Role of gene duplication in evolution. Genome. 1989; 31:304–10. [PubMed: 2687099]

Pai CY, Kuo TS, Jaw TJ, Kurant E, Chen CT, Bessarab DA, Salzberg A, Sun YH. The Homothorax homeoprotein activates the nuclear localization of another homeoprotein, extradenticle, and suppresses eye development in Drosophila. Genes Dev. 1998; 12:435–46. [PubMed: 9450936]

Pan D, Rubin GM. Targeted expression of teashirt induces ectopic eyes in Drosophila. Proc Natl Acad Sci U S A. 1998; 95:15508–12. [PubMed: 9860999]

Pignoni F, Hu B, Zavitz KH, Xiao J, Garrity PA, Zipursky SL. The eye-specification proteins So and Eya form a complex and regulate multiple steps in Drosophila eye development. Cell. 1997; 91:881–91. [PubMed: 9428512]

Punzo C, Plaza S, Seimiya M, Schnupf P, Kurata S, Jaeger J, Gehring WJ. Functional divergence between eyeless and twin of eyeless in Drosophila melanogaster. Development. 2004; 131:3943–53. [PubMed: 15253940]

Quiring R, Walldorf U, Kloter U, Gehring WJ. Homology of the eyeless gene of Drosophila to the Small eye gene in mice and Aniridia in humans. Science. 1994; 265:785–9. [PubMed: 7914031]

Relaix F, Buckingham M. From insect eye to vertebrate muscle: redeployment of a regulatory network. Genes Dev. 1999; 13:3171–8. [PubMed: 10617565]

Rudel D, Sommer RJ. The evolution of developmental mechanisms. Dev Biol. 2003; 264:15–37. [PubMed: 14623229]

Salzer CL, Kumar JP. Identification of retinal transformation hot spots in developing Drosophila epithelia. PLoS One. 5:e8510. [PubMed: 20062803]

Scott K, Brady R Jr. Cravchik A, Morozov P, Rzhetsky A, Zuker C, Axel R. A chemosensory gene family encoding candidate gustatory and olfactory receptors in Drosophila. Cell. 2001; 104:661–73. [PubMed: 11257221]

Seimiya M, Gehring WJ. The Drosophila homeobox gene optix is capable of inducing ectopic eyes by an eyeless-independent mechanism. Development. 2000; 127:1879–86. [PubMed: 10751176]

Serikaku MA, O'Tousa JE. sine oculis is a homeobox gene required for Drosophila visual system development. Genetics. 1994; 138:1137–50. [PubMed: 7896096]

Shen W, Mardon G. Ectopic eye development in Drosophila induced by directed dachshund expression. Development. 1997; 124:45–52. [PubMed: 9006066]

Shimeld SM. Gene function, gene networks and the fate of duplicated genes. Semin Cell Dev Biol. 1999; 10:549–53. [PubMed: 10597639]

Shippy TD, Tomoyasu Y, Nie W, Brown SJ, Denell RE. Do teashirt family genes specify trunk identity? Insights from the single tiptop/teashirt homolog of Tribolium castaneum. Dev Genes Evol. 2008; 218:141–52. [PubMed: 18392876]

Silver SJ, Rebay I. Signaling circuitries in development: insights from the retinal determination gene network. Development. 2005; 132:3–13. [PubMed: 15590745]

Simpson TI, Price DJ. Pax6; a pleiotropic player in development. Bioessays. 2002; 24:1041–51. [PubMed: 12386935]

Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 2007; 24:1596–9. [PubMed: 17488738]

Teichmann SA, Babu MM. Gene regulatory network growth by duplication. Nat Genet. 2004; 36:492–6. [PubMed: 15107850]

Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22:4673–80. [PubMed: 7984417]

Treisman JE. A conserved blueprint for the eye? Bioessays. 1999; 21:843–850. [PubMed: 10497334]

Vanin EF. Processed pseudogenes: characteristics and evolution. Annu Rev Genet. 1985; 19:253–72. [PubMed: 3909943]

Wagner A. Genetic redundancy caused by gene duplications and its evolution in networks of transcriptional regulators. Biol Cybern. 1996; 74:557–67. [PubMed: 8672563]

Wawersik S, Maas RL. Vertebrate eye development as modeled in Drosophila. Hum Mol Genet. 2000; 9:917–25. [PubMed: 10767315]

Weasner B, Salzer C, Kumar JP. Sine oculis, a member of the SIX family of transcription factors, directs eye formation. Dev Biol. 2007; 303:756–71. [PubMed: 17137572]

Weasner BM, Weasner B, Deyoung SM, Michaels SD, Kumar JP. Transcriptional activities of the Pax6 gene eyeless regulate tissue specificity of ectopic eye formation in Drosophila. Dev Biol. 2009; 334:492–502. [PubMed: 19406113]

Weasner BP, Kumar JP. The non-conserved C-terminal segments of Sine Oculis Homeobox (SIX) proteins confer functional specificity. Genesis. 2009; 47:514–23. [PubMed: 19422020]

Wilgenbusch JC, Swofford D. Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics. 2003 Chapter 6, Unit 6 4.

Xu PX, Zheng W, Huang L, Maire P, Laclef C, Silvius D. Six1 is required for the early organogenesis of mammalian kidney. Development. 2003; 130:3085–94. [PubMed: 12783782]

Yao JG, Sun YH. Eyg and Ey Pax proteins act by distinct transcriptional mechanisms in Drosophila development. EMBO J. 2005; 24:2602–12. [PubMed: 15973436]

Yao JG, Weasner BM, Wang LH, Jang CC, Weasner B, Tang CY, Salzer CL, Chen CH, Hay B, Sun YH, Kumar JP. Differential requirements for the Pax6(5a) genes eyegone and twin of eyegone during eye development in Drosophila. Dev Biol. 2008; 315:535–51. [PubMed: 18275947]

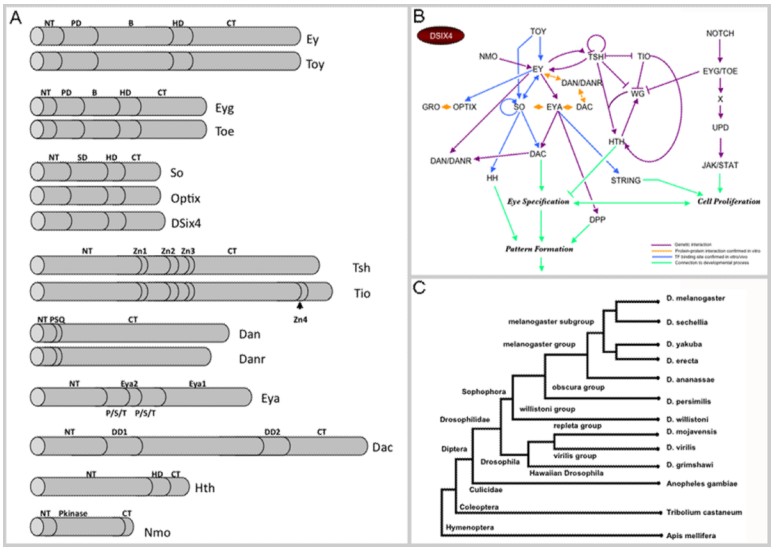

Figure 1. Schematic of Protein Structures, the Retinal Determination Network and a Drosophila Species Tree

(A) RD proteins. NT: N terminal, CT: C terminal, B: Central linker, PD: Paired domain, HD: Homeodomain, SD: SIX domain, Zn(1-4): Zn finger (1-4), PSQ: Pipsqueak domain, P/S/T: Pst domain, Eya1 and Eya2: Eya 1 and Eya2 Domain; DD1 and DD2; Dac1 and Dac2 domains (B) RD network. Purple connectors indicate genetic interactions; Orange arrows indicate confirmed protein-protein interactions; Blue arrows indicate confirmed transcription factor binding; Green arrows show connections to developmental processes. The X downstream of the Notch pathway shows an as yet unidentified molecule. (C) Phylogenetic tree of the species used in this study.

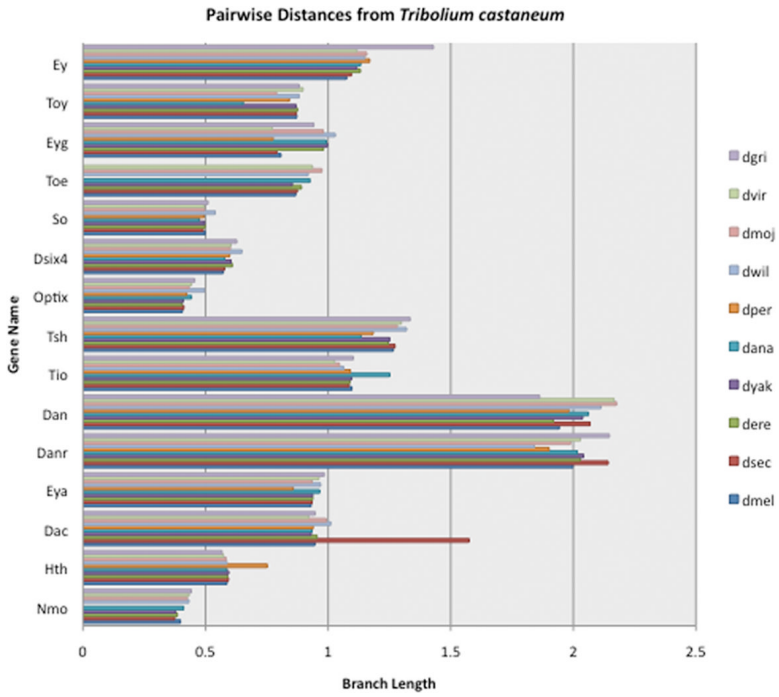

Figure 2. Divergence of Retinal Determination Genes within Drosophila

Branch lengths were calculated using distance-based trees generated from *Drosophilid* nucleotide sequences using *Tribolium castaneum* as most recent common ancestor (MRCA). While all the genes have different divergence rates, the genes are not evolving rapidly in any particular lineage.

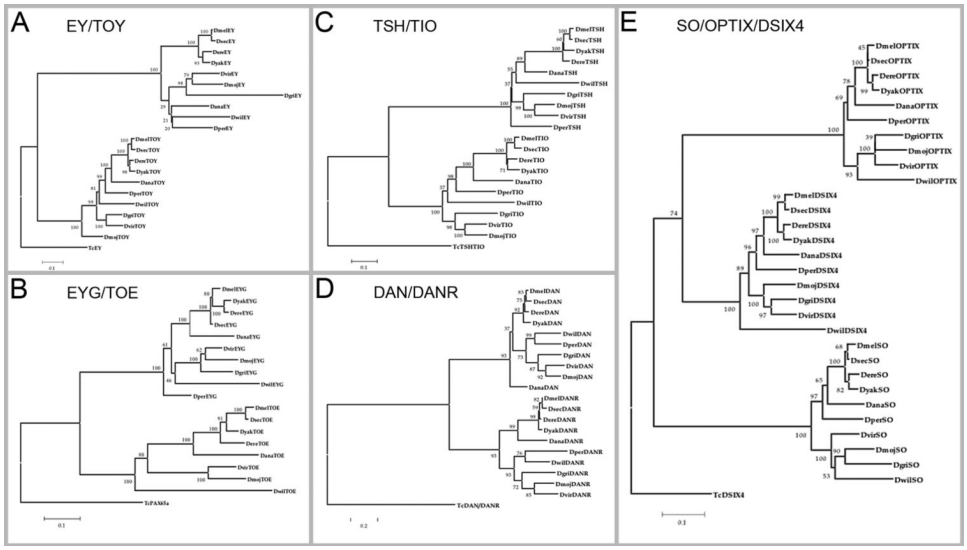

Figure 3. Phylogenetic Analysis of Paralog Pairs within the Retinal Determination Network

(A) Ey/Toy. (B) Eyg/Toe. (C) Tsh/Tio. (D) Dan/Danr. (E) So/Optix/DSix4. All duplicate genes are evolving at significantly different rates compared to their sister gene, except for Eyg/Toe and Dan/Danr.

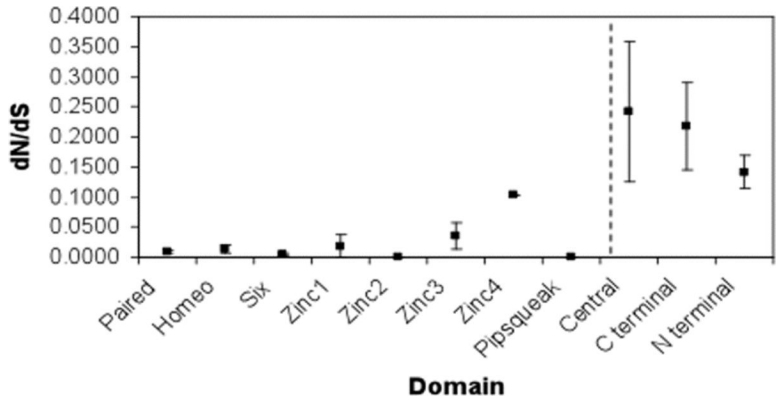

Figure 4. Selective Pressures on Functional Domains and Non-Conserved Regions

The highly structured regions of the genes in RD cascade (Paired, Homeo, Paired, Zinc1-4, Pipsqueak) are under significantly higher purifying selection than the non-conserved regions (N-terminal, B linker and C- terminal). The error bars indicate standard errors. This is likely to allow the genes to remain connected in a highly regulated network through the DNA-binding and protein-interaction domains, while the N-terminal, B linker and C- terminal segments accumulate mutations and gain new functions.

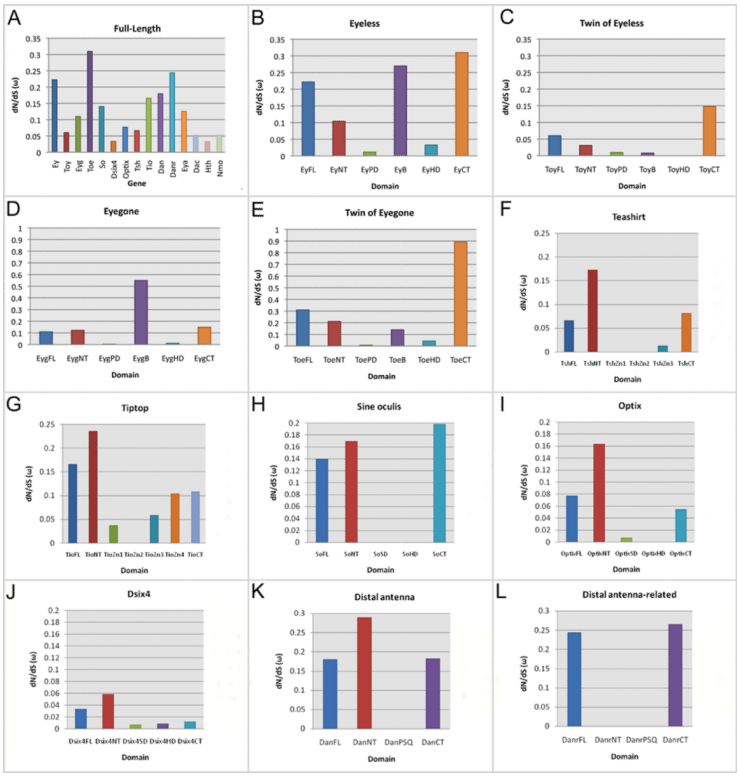

Figure 5. Selection Signatures of Functional Domains and Non-Conserved Regions within the Retinal Determination Network

(A) Full length genes. (B,C) Ey/Toy. (D,E) Eyg/Toe. (F,G) Tsh/Tio. (H-J) So/Optix/DSix4 (K,L) Dan/Danr. All genes are under varying degrees of purifying selection. The duplicate genes have different patterns of selection across their coding regions, with the highly structured regions being more constrained than the non-structured regions.

Table 1

d_n/d_s Raw Values (full-length, functional domains, non-conserved segments)

The raw d_n/d_s values obtained for the full-length genes within the melanogaster group as well as those obtained for individual functional domains and non-conserved regions.

	FL	HD	PD	SD	Znl	Zn2	Zn3	Zn4	PSQ	NT	B	CT
Ey	0.222399	0.32549	0.011792	-	-	-	-	-	-	0.103674	0.270092	0.310097
Toy	0.0604	0	0.009882	-	-	-	-	-	-	0.031145	0.08752	0.147523
Eyg	0.109955	0.012442	0.00286	-	-	-	-	-	-	0.123141	0.055385	0.149491
Toe	0.309762	0.044019	0.008524	-	-	-	-	-	-	0.211779	0.139744	0.892635
So	0.140439	0	-	0	-	-	-	-	-	0.169416	-	0.197658
DSix4	0.033421	0.008052	-	0.006466	-	-	-	-	-	0.057762	-	0.0115
Optix	0.077049	0	-	0.00717	-	-	-	-	-	0.162835	-	0.054253
Tsh	0.066363	-	-	-	0	0	0.012557	-	-	0.1726	-	0.081748
Tio	0.166197	-	-	-	0.036876	0	0.057631	0.1043	-	0.235343	-	0.10774
Dan	0.179568	-	-	-	-	-	-	-	0	0.289031	-	0.182197
Danr	0.243908	-	-	-	-	-	-	-	0	0	-	0.26496
Eya	0.125578	-	-	-	-	-	-	-	-	0.22408	-	0.042771
Dac	0.051559	-	-	-	-	-	-	-	-	0.025008	-	0.09518
Hth	0.032541	-	-	-	-	-	-	-	-	0.003011	-	-
Nmo	0.0117	-	-	-	-	-	-	-	-	0.42487	-	0.123648

Table 2
Number of clade-specific and group-specific changes in residues

The paralog pairs show some number of residue changes even within the highly structured domains. This is likely to account for the changes in gene regulation and protein binding that we see between the paralog pairs.

Gene Pair	Domain	Clade Specific Residues	Group Specific Residues
Ey/Toy	PD	14	0
Ey/Toy	HD	0	3
Eyg/Toe	PD	3	1
Eyg/Toe	HD	5	3
Tsh/Tio	Zn1	4	3
Tsh/Tio	Zn2	4	0
Tsh/Tio	Zn3	4	3
So/Optix/DSix4	HD	28	0
So/Optix/DSix4	SIX	61	1
Dan/Danr	PSQ	0	0