Universidade de Lisboa

Faculdade de Ciências

Departamento de Biologia Animal

Gene function diversification upon duplication

Luís Miguel Baudouin Gonzalez

Dissertação
Mestrado em Biologia Evolutiva e do Desenvolvimento
2013

Universidade de Lisboa

Faculdade de Ciências

Departamento de Biologia Animal

Gene function diversification upon duplication

Luís Miguel Baudouin Gonzalez

Dissertação
Mestrado em Biologia Evolutiva e do Desenvolvimento

Orientador interno: Doutor Élio Sucena
Orientador externo: Kohtaro Tanaka
2013

Acknowledgments

I would like to thank Élio Sucena for the opportunity of working in such an interesting topic, for the few but important discussions about the work I was developing, and for the last hour revision of the thesis.

Kohtaro Tanaka for all the help he gave me during this year, both in methodological and theoretical aspects. This work would not be possible without his supervision.

Diogo Manoel for all the help in the acquisition of the transgenic fly lines, and for the extra rows of embryos he aligned for me.

The other members of the Evo Devo lab, Nelson, Alexandre, Vítor, Barbara and Rui, who all helped me at some point and made me feel integrated.

All the members of the labs in Vasco da Gama wing of IGC, for creating such an excellent ambient to work.

To my family who always supported my choices.

Abstract

Gene duplication facilitates the evolution of new gene functions by relaxing evolutionary pressure on the duplicates. Besides pseudogenization, three outcomes can occur: both copies keep the ancestral function (e.g., Positive dosage); one maintains the ancestral function and the other gains a new function (Neofunctionalization); ancestral functions are partitioned between the copies (Subfunctionalization). *Cis*-regulatory evolution, mainly enhancers, has gained weight in explaining this process, due to its lesser pleiotropic nature relative to protein evolution. However, knowledge on the *cis*-regulatory changes responsible for duplicate functional diversification is still lacking. In our lab, we use clusters III and V of the Ly6 gene family of *Drosophila melanogaster*, which underwent sequential duplications during the dipteran evolution, as a gene duplication evolutionary model. At each duplication event, one of the duplicates maintained the ancestral expression pattern while the other acquired novel expression domains (Neofunctionalization). *CG9336* and *CG9338*, found in *D. melanogaster*, are the most recent duplicates, having already divergent expression patterns. An interesting fact is that expression divergence is not only seen between these duplicates and the unduplicated orthologue found in *Ceratitis capitata*, but also between duplicates in different species of Drosophilids. Given the high frequency of duplication events within a short time-scale observed in these genes, they are a good model to study *cis*-regulatory evolution.

In this study, I focus on the most recent duplication, *D. melanogaster CG9336* and *CG9338*, to examine the genomic location of the enhancers responsible for the expression pattern divergence observed. Intergenic and Intronic sequences of these genes were tested on a Gal4 reporter vector for enhancer activity. All regions tested showed enhancer activity and were able to recapitulate most of the duplicates expression patterns, both new and conserved. Sequence conservation analysis was also performed for the three Drosophilid species that showed expression pattern divergence.

Keywords: Enhancer, *CG9336*, *CG9338*, Expression pattern, *Cis*-regulatory evolution.

Resumo

A duplicação de genes é um processo molecular que facilita a evolução de novos genes com novas funções. Isto acontece porque a presença de uma segunda cópia relaxa a pressão selectiva sobre estes genes, permintindo que se acumulem mutações nestes. A pseudogenização de uma das cópias é o destino mais provável após um evento de duplicação, assumindo que a fixação dá-se através de selecção neutral (o que é normalmente assumido), mas outros resultados podem ocorrer: ambas as cópias podem manter a função ancestral (e.g., dosagem positiva ou robustez genética); uma cópia mantém a função ancestral enquanto a outra evolui adquirindo uma nova função (modelo de neofuncionalização); ou as funções ancestrais podem ser repartidas por ambas as cópias (modelo de subfuncionalização). Existem vários exemplos que ligam divergência fenotípica à origem de novos genes através de duplicação génica, mas os processos evolutivos por detrás deste tipo de evolução ainda não são inteiramente percebidos.

A evolução *cis*-regulatória, principalmente a evolução de enhancers (uma sequência de DNA não codificante que possui locais de ligação a factores de transcrição, induzindo a expressão do gene a que está associada), tem ganho importância em explicar o processo por detrás da evolução de genes duplicados, devido ao facto de que modificações na sequência codificante de um gene são de uma natureza pleitrópica maior quando comparadas com modificações em sequências regulatórias (i.e., têm uma probabilidade maior de afectarem a função do gene, pois podem modificar a proteína codificada por este gene, do que modificações num enhancer, que só afecta parte da expressão do gene). Os enhancers podem sofrer vários tipos de modificações que podem causar mudanças de expressão no gene regulado. Eles próprios podem acumular mutações, o que por sua vez pode causar 1) a perda desse mesmo enhancer, sendo perdida a expressão do gene por ele regulada, ou 2) a modificação dos locais de ligação a factores de transcrição, mudando o seu domínio de expressão ou criando um novo para além do já existente. Modificações na localização de um enhancer podem também causar mudanças na expressão do gene por ele regulada.

Assim, os padrões de expressão de duas cópias de um gene podem ser mudados através da modificação de enhancers, de modo a que estes sejam expressos em tecidos diferentes, deixando assim de ser redundantes e

podendo evoluir novas funções ou dividir as funções ancestrais. Existem vários trabalhos que ligam mudanças *cis*-regulatórias à evolução de novas funções, mas os mecanismos evolutivos pelos quais estas novas funções foram adquiridas estão pouco estudados e fracamente compreendidos. No nosso laboratório, é usado como modelo de evolução da duplicação de genes os clusters III e V da família de genes Ly6 de *Drosophila melanogaster*, os quais sofreram duplicações sequenciais durante a evolução dos Dípteros. Os clusters são compostos por 9 genes, dos quais 8 surgiram através de sucessivas duplicações. O padrão de expressão ancestral é mantido por uma das cópias em cada evento de duplicação, enquanto a outra cópia diverge e evolui novos padrões de expressão.

Os genes *CG9336* e *CG9338* do cluster V de *D. melanogaster* resultam da duplicação mais recente, a qual se deu após a separação entre os Tephritidae e os Drosophilidae. Um ortólogo do gene ancestral pode ser encontrado na espécie *Ceratitis capitata*. O padrão de expressão ancestral, caracterizado pela expressão no sistema nervoso e epiderme do embrião, é mantido quase únicamente pelo gene *CG9336*, enquanto que o gene *CG9338* possui uma nova expressão nos hemócitos. No entanto, a expressão na epiderme parece ter sido quase perdida e um novo domínio de expressão é visível no tubo cardíaco do embrião no gene *CG9336*. Outro facto interessante é a fraca expressão do gene *CG9338* no sistema nervoso. O padrão de expressão destes genes diverge também entre diferentes espécies de Drosophilídios, nomeadamente as expressões na epiderme e no tubo cardíaco. A expressão na epiderme parece ser gradualmente perdida e a expressão no tubo cardíaco parece ser gradualmente ganha. No entanto, existem domínios de expressão que parecem manter-se conservados durante a evolução destes genes (e.g., expressão no sistema nervoso), o que suscita a ideia de conservação de enhancers.

A divergência observada nestes genes deu-se numa pequena escala de tempo, tornando estes genes um bom modelo para estudar a evolução *cis*-regulatória e a sua influência na evolução de genes duplicados. Neste estudo, eu foco-me na duplicação mais recente, os genes *CG9336* e *CG9338* de *D. melanogaster*, com o objectivo de examinar a localização genómica dos enhancers responsáveis pela divergência de padrões de expressão observada. Para este efeito, eu isolei, clonei, e sequenciei para verificação de erros produzidas por PCR, as sequências dos intrões e das regiões intergénicas 5′ e 3′ dos genes *CG9336* e *CG9338*, usando-os para construir vectores repórter Gal4 para testar se estes possuem actividade de enhancers. Estes vectores foram depois transformados em *D. melanogaster*, adquirindo linhas transformantes para cada uma destas regiões e cruzando-as com uma linha UAS-GFP para verificar a expressão do gene repórter. Análise de conservação de sequência, possível através da ferramenta bioinformática mVista, foi feita para as três espécies de Drosophilídios (*D. ananassae*, *D. melanogaster* e *D. virilis*) com padrões de expressão divergentes, para verificar se existem padrões de expressão que possam ser relacionados com os resultados adquiridos com os gene repórter, e se existem mudanças nestes padrões entre estas espécies que possam explicar as diferenças observadas, pois é esperado que as sequências de enhancers sejam mais conservadas do que outras sequências não-codificantes.

Todas as regiões testadas mostraram actividade de enhancer, sendo capazes de recriar quase todos os domínios de expressão dos genes estudados, tanto os novos como os conservados. Regiões que regulam expressão na epiderme, no sistema nervoso, no tubo cardíaco e nos hemócitos foram encontradas. Ainda mais, picos de conservação resultantes da análise em mVista desta região genómica são encontrados em cada uma destas regiões, possívelmente representando os enhancers aqui encontrados. Diferenças na conservação destes picos entre espécies suscitam modificações na sequência destes enhancers, significando que estes sofreram mudanças durante a evolução destes genes. Estas mudanças podem ser explicativas da forma como os padrões de expressão divergiram.

Parece existir um padrão de duplicação, seguida de modificação de enhancers, a qual pode explicar como estes genes foram mantidos em termos de evolução de genes duplicados. Porém, a evolução dos módulas *cis*-regulatórios destes genes parece longe de ser linear, sendo mais complexa do que era esperado, e os dados deste trabalho não são suficientes por si só para explicar a extensão das modificações por detrás da divergencia observada, pois apenas nos dão informações sobre um único estado evolutivo. Para uma melhor compreensão da evolução *cis*-regulatória por detrás das mudanças de expressão observadas nestes genes entre Drosophilídeos e o gene encontrado em *C. capitata*, o tipo de estudo aqui feito deve ser estendido às outras

espécies que apresentam diferenças nos padrões de expressão destes genes. Progressos nesta direcção foram já feitos por mim, através do isolamento e clonagem de algumas destas regiões noutras espécies de Drosophilídeos que serão utilizadas para criar mais genes repórter para testar a capacidade destas regiões de activar expressão. Estes resultados serão depois comparados com os resultados deste trabalho para elucidar com é que os enhancers responsáveis pelos vários domínios de expressão dos genes duplicados evoluíram para dar origem às diferenças de expressão observadas.

Palavras-chave: Enhancer, *CG9336*, *CG9338*, Padrão de expressão, Evolução *cis*-regulatória.

Index

I. Introduction

II. Materials and Methods

III. Results

IV. Discussion

I. Introduction

1. Gene duplication and genetic innovation

The origin and evolution of new genes with novel functions has been a much-discussed matter, particularly with regards to their contribution to phenotypic divergence. Being a common phenomenon in all Eukaryotes, gene duplication is now thought to be the primary driving force behind the origin of new genes [1]. This was first theorized by Susumu Ohno [2], who suggested that the presence of a second copy of a gene would relax selection on the duplicates, allowing one of them to diverge functionally while the other maintains the ancestral function (neofunctionalization) Other possible outcomes are the partitioning of the ancestral gene functions between the two copies (subfunctionalization) or preservation by selection for gene dosage. Yet, loss of one copy by pseudogenization is the most probable fate [1].

Since Ohno's work, many examples of functional divergence after gene duplication have been found [5-8]. Notwithstanding, most of our knowledge on this phenomenon resides at the phylogenetic, functional and genomic levels, whereas insights on the evolutionary mechanisms behind maintenance and evolution of gene duplicates are poorly understood and mostly theoretical [4].

1.1. Which mechanisms underlie gene duplication?

There are four main mechanisms by which a gene can be duplicated: transposition, retrotransposition, segmental duplication, and whole genome duplication (Polyploidization) [3]. Due to the "cut-and-paste" process of transposition, duplication is only likely when associated with lateral gene transfer, frequent only in bacteria and archaebacteria. Retrotransposition is frequent in many eukaryotes, and consists in the integration of a reverse transcribed mRNA into a random site in the genome, giving rise to an intron-less gene with a poly-A tail (retrogene). Due to the lack of regulatory sequences, retrogenes rarely give rise to new genes.

The duplication of a chromosomal segment containing whole genes or gene fragments is called Segmental duplication (or Tandem duplication; Fig. 1, top). This gives rise to an identical segment tandem to the duplicated one. This process is commonly the result of unequal crossing over, but it can also be the result of other non-homologous mechanisms [3]. Polyploidization events are recurrent in eukaryotes and give rise to a whole new set of duplicated genes that can be tinkered with (Fig. 1, bottom). This creates a huge opportunity for gene function divergence, which could cause a major burst of genomic complexity. However, genomic transmission from one generation to the other is an obstacle to polyploidy preservation, making this a rare event.

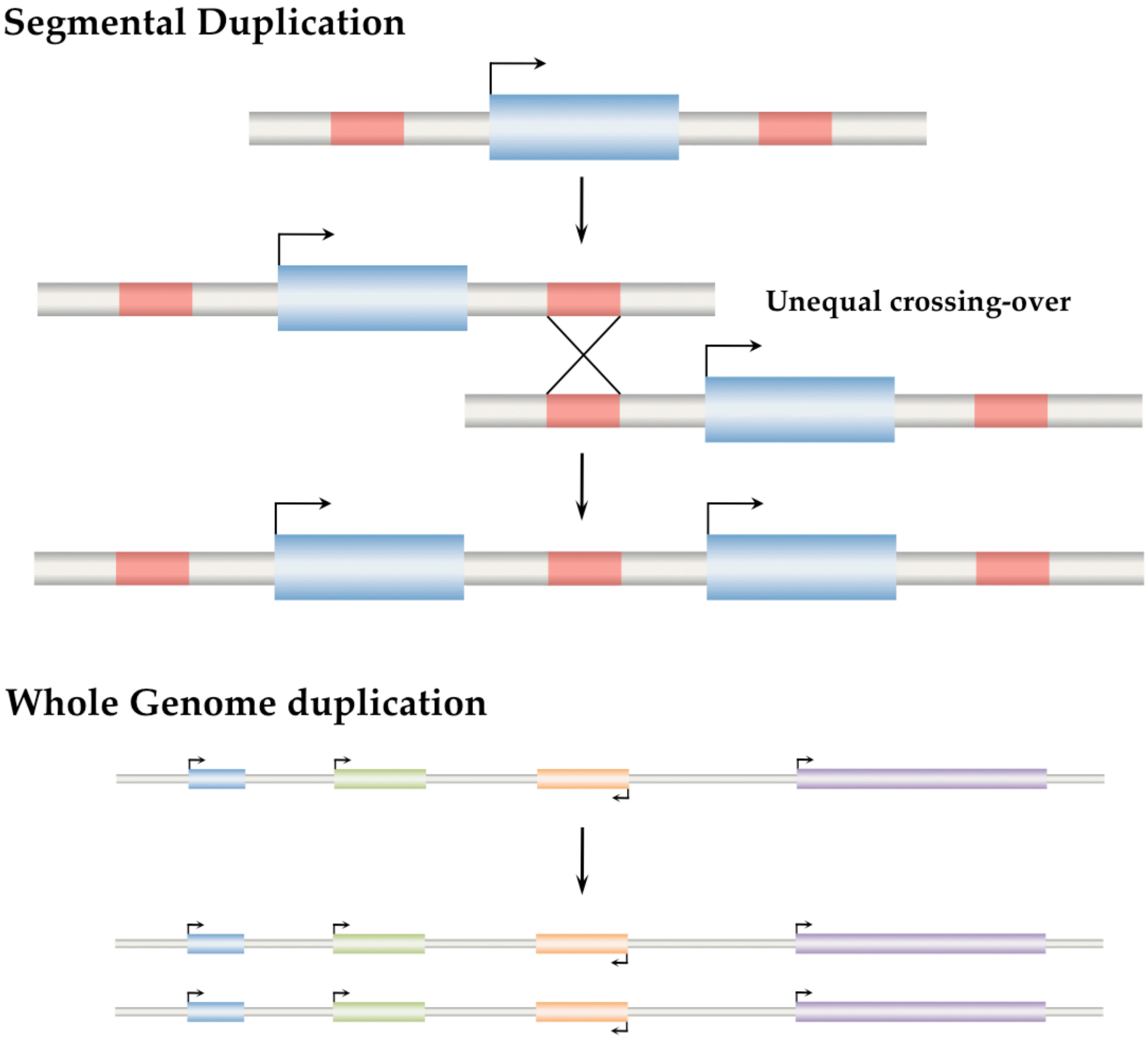

Figure 1 – Mechanisms of gene duplication. Schematics depicting the most common mechanisms behind gene duplication: segmental duplication and whole genome duplication. In the first case, unequal crossing-over between similar sequences (red boxes) gives rise to a second copy of the gene. In the second, abnormal cell division (commonly in meiosis) give rise to an organism with a complete set of duplicated chromosomes, duplicating every single gene.

1.2. How are duplicates fixed, maintained and preserved?

In order for a duplicate to be preserved in a genome, it has first to be fixed in the population and then maintained, so that each copy can diverge and be selected [2, 4]. Many gene duplication evolutionary models and hypothesis have been proposed, but their relative importance and applicability is unknown [4]. Three of these models are of particular interest to this study, as they are the only ones in which divergence between duplicates occur after fixation. These models are neofunctionalization, subfunctionalization or duplication-degeneration-complementation (DDC), and specialization. All three models assume that the duplicates are fixed by drift in the population, varying only in the evolutionary mechanisms by which they diverge and are preserved [4].

1.2.1. Neofunctionalization. First proposed by Ohno [2], this model states that the extra gene copy is functionally redundant, therefore, free from selective pressure. The model assumes that the original copy is sufficient to carry out the function of the ancestral gene, remaining under purifying selection, freeing the extra copy to diverge by neutral selection. Even though pseudogenization of the new copy by accumulation of neutral loss-of-function mutations is the most probable outcome, Ohno hypothesized that, in some cases, the acquisition of a new function by the emergence of fortuitous gain-of-function mutations could occur. In this occasion, the duplicate is no longer redundant and is preserved by positive selection, opening the possibility for further functional improvements by accumulation of new advantageous mutations. However, this has to take place in a relatively short time period, since pseudogenization is quick under neutral selection.

1.2.2. Duplication-degeneration-complementation (DDC). Also based on the redundancy of gene duplication, the DDC or subfunctionalization model, first suggested by Force and Lynch [9, 10], changes only one of the assumptions of the neofunctionalization model: selection is relaxed in both copies, instead of just one. This makes it possible that different degenerative mutations accumulate on each copy, causing each copy to be unable to fully perform the ancestral function alone. The most probable outcome is still the pseudogenization of one copy, though, if each duplicate acquires different degenerative mutations (i.e., mutations affecting different aspects of gene function), one can compensate the other, becoming subfunctionalized. So, the original function gets partitioned between the two duplicates, forcing both of them to be maintained. The combined functions and fitness contributions of both paralogues should be equivalent to the ones of the unduplicated orthologue. In some cases, these subfunctionalized duplicates can still go through further changes to achieve optimum protein function.

1.2.3. Specialization. The specialization or escape from adaptive conflict (EAC) model, postulated by Hughes [11], has the same base assumptions of the previous models, fixation by drift and duplicate redundancy. The new idea behind this model is that the unduplicated gene performs two or more functions that cannot be independently improved. This creates a scenario where the presence of an extra copy opens the possibility for specialization and improvement of the original functions. In this case, gene duplication evolution is driven by positive selection to specialize, sometimes leading to a similar outcome to the DDC model. This model also predicts that paralogues fitness contribution is greater than that of the unduplicated orthologue.

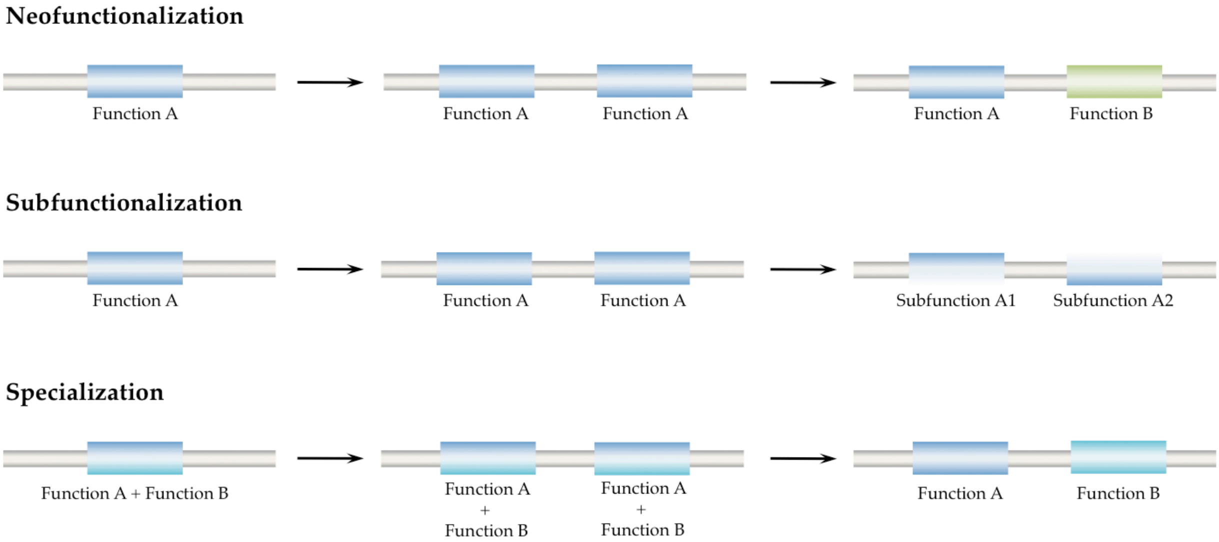

Figure 2 – Models of gene duplication evolution. In the neofunctionalization model, a gene with function A is duplicated, and one of the duplicates acquires a new function B. In the subfunctionalization model, a gene with function A is duplicated, and the ancestral function is divided between the duplicates (A1 and A2). In the specialization model, a gene with functions A and B is duplicated, and the ancestral functions are divided between the duplicates.

2. Cis-regulatory evolution

Over the last decade, a rapidly growing number of case studies were able to correlate *cis*-regulatory changes to phenotypic divergence [13-16]. For this reason, *cis*-regulatory evolution can be of great relevance to gene duplication evolution. The idea that mutations affecting gene expression regulation could be the main source of phenotypic evolution arose from the fact that many homologous genes, with conserved protein functions between species, show differences in their expression that led to phenotypic divergence. This contributed to the reasoning that mutations in the coding region have a higher probability of being deleterious than mutations in tissue-specific *cis*-regulatory elements (CREs), because changes in coding sequence are expected to be more pleitropic (i.e., they affect more gene functions) than changes in *cis*-regulatory sequences [12]. The line of thought that gave rise to this prediction was that a change in protein function probably affects all domains where the gene is active, while a change in a CRE should only affect the domain it regulates.

The type of CRE more commonly thought to underlie expression pattern divergence is the enhancer, due to its higher variability between species [17]. An enhancer is composed of several transcription factor (TF) binding sites, which are needed to regulate specific spatial and temporal expression domains of the gene they are associated to. Their genomic location is often conserved between species [18] and it has been shown that a small number of mutations can change their *cis*-regulatory functions [13]. As a result of this modular organization, a mutation in a single enhancer, although it can change the gene expression pattern, is expected to be of smaller effect and functionally independent of other enhancers, making them good candidates for expression pattern divergence [12].

2.1. Which cis-regulatory modifications underlie expression pattern divergence?

The types of mutations that are most commonly found to underlie *cis*-regulatory changes are substitutions, deletions and insertions [12]. These can alter the TF binding sites of an enhancer, by modifying its affinity, creating new or disrupting existing ones. Changes in spacing between TF binding sites can also modify enhancer activity [12].

Enhancers can be created *de novo* in a sequence previously lacking *cis*-regulatory activity, by accumulating mutations that give rise to new TF binding sites (Fig. 3, top right corner) [19]. If this new enhancer is close enough to a gene so that it can regulate its activity, a new expression domain arises. Contrary to this, accumulation of disruptive mutations in an enhancer can cause it to be no longer functional, resulting in the loss of the respective expression domain (Fig. 3, bottom right corner). Another possible process, called co-option, involves the modification of an existing enhancer to drive an additional expression pattern (Fig. 3, bottom left corner) [20]. In this case, mutations that create new binding sites and/or change the organization of the existing ones within an enhancer, result in a novel expression domain. This is thought to be the most common process behind the appearance of novel expression patterns [12].

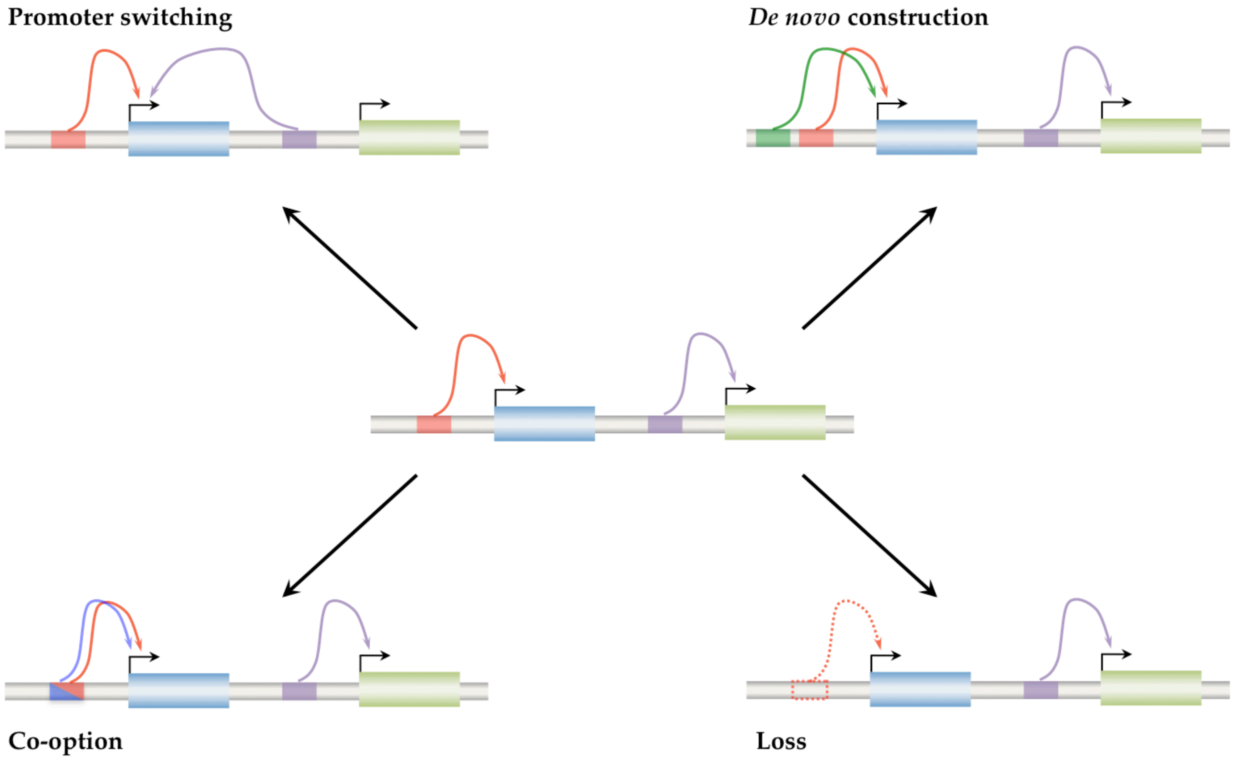

Figure 3 – Enhancer modifications underlying expression pattern divergence. Four enhancer modifications that can change expression of the regulated gene are shown. A previously existing enhancer can be modified to express in novel tissues (co-option) or simply switch which promoter it interacts with (promoter switching). *De novo* creation of an enhancer in a region previously lacking *cis*- regulatory activity or the loss of previously existing one is also possible [12].

Transposable elements can also generate *cis*-regulatory changes by transporting an enhancer into another position, adding an expression domain to close by genes. They can also cause expression divergence by increasing or decreasing (e.g. in the previous location of a cut and paste element) the relative distance of enhancers to nearby genes, possibly changing which one they regulate (promoter switching [21]; Fig. 3, top left corner).

2.2. *How can cis-regulatory changes shape gene duplication evolution?*

The importance of *cis*-regulatory sequences as an essential component of gene duplication evolution was first suggested by Force *et al*. [9], using this idea as an explanation to why so many duplicates show partitioning of the ancestral expression pattern. They propose that mutations in regulatory sequences rather than mutations in coding sequences underlie divergence of expression patterns and consequent maintenance of the duplicates. Assuming that, in the duplication event, the regulatory elements are duplicated along with the gene they regulate, these enhancers become as redundant as the duplicates. Now under relaxed selection, these enhancers can be modified, giving rise to new expression domains, or be lost through accumulation of mutations [9, 10].

This hypothesis can be added to the three models of gene duplication evolution described above. If a duplicated enhancer acquires mutations that modify the expression pattern it regulates (i.e., it was co-opted), then the duplicated gene linked to this enhancer becomes neofunctionalized, while the other retains the ancestral functions. Furthermore, any of the copies can undergo this process if we assume both are under relaxed purifying selection. Other possibility is the *de novo* formation of an enhancer controlling one of the duplicates.

In the case of the DDC model, subfunctionalization can be the result of a loss of different enhancers for each duplicate by accumulation of degenerative mutations, causing the ancestral expression pattern to be partitioned between the copies. In this scenario, each copy has to compensate for the lack of expression of the other. EAC can be similarly explained: if the functions differ in their spatial and/or temporal expression domains, then specialization can be achieved by selection for partition of specific sets of enhancers between duplicates responsible for each function.

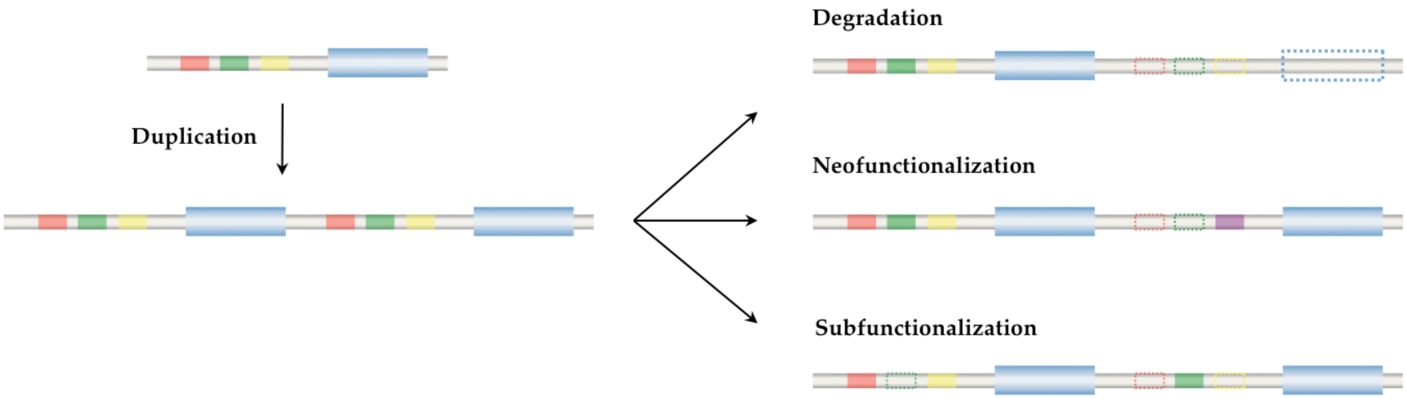

Figure 4 – Gene duplication evolution through *cis*-regulatory changes. To simplify the diagram, an assumption that all the ancestral enhancers are also duplicated is made. As depicted in the diagram, Neofunctionalization can be achieved by the modification (e.g., co-option) of a duplicated enhancer regulating one gene copy and maintenance of the ancestral enhancers on the other gene copy. Similarly, Subfunctionalization can be achieved by loss of different enhancers regulating each gene duplicate.

3. D. melanogaster *clusters III and V Ly6 genes: a model of gene duplication evolution*

Members of the Ly6 superfamily are widely distributed in the metazoan genomes and are characterized by proteins with one or more three-finger-domains, which are composed of a single polypeptide chain of 60-100 amino acid residues, folded into three adjacent loops or fingers linked to a hydrophobic palm [22]. Only 8 or 10 cysteine residues in stereotyped positions seem to be conserved between Ly6 primary sequences, but their three-dimensional structures are very similar [22]. Involved exclusively in protein-protein interactions, Ly6 protein functions range from receptors (e.g. TGFβ receptors) to soluble ligands (e.g. neurotoxins of elapid snakes).

There are 45 members of the Ly6 superfamily in *Drosophila melanogaster*, many of them grouped into six clusters [23]. Only seven of these were functionally characterized: *boudin*, *crooked*, *crimpled*, and *coiled* are all required for septate junction formation, *retroactive* is needed for cuticle organization in the larva, *sleepless* is required for sleep cycle regulation and *twit* is responsible for neurotransmitter release in the neuromuscular junction [23, 24, 25]. Lineage-specific expansion of the Ly6 gene family can be observed in insects, especially in the higher diptera. Only 14 Ly6 genes have been identified in *Apis mellifera*, most of which have fly orthologues, suggesting that Ly6 gene family has undergone sudden events of duplication, followed by rapid functional diversification, across the higher diptera. This makes the insect Ly6 family a good model for understanding the origin of novel functions in new genes.

A subset of this family, namely the Ly6 genes in clusters III and V of *D. melanogaster*, has been the focus of a recent study in our lab (unpublished data). This subset is composed of 9 genes, 3 form cluster III (*atilla*, *crooked* and *CG6583*) and 6 form cluster V (*CG31676*, *twit*, *CG9336*, *CG9338*, *CG31675* and *CG14401*). In the study, it was found that 8 of these genes arose from sequential duplications of a single ancestral gene (Fig. 5). Phylogenetic analysis placed the first duplication event before the split of Hymenoptera from the rest of the Holometabola, in which the *crooked*-like ancestral gives rise to *crooked* and the ancestral gene of cluster V (a-cV). The latter underwent two duplication events, one after the split of lower and higher diptera, and another one before the split of Phoridae (*Megaselia*), giving rise to *CG31676*, *twit*, the ancestral of 9336/38 lineage (a-36/38) and *attila*. Two more duplication events after the split of Phoridae gave rise to *CG31675*, *CG14401* and the ancestral gene 36/38. *CG9336* and *CG9338* are the most recent duplicates, generated through a final duplication event after the split of Tephritidae (*Ceratitis*).

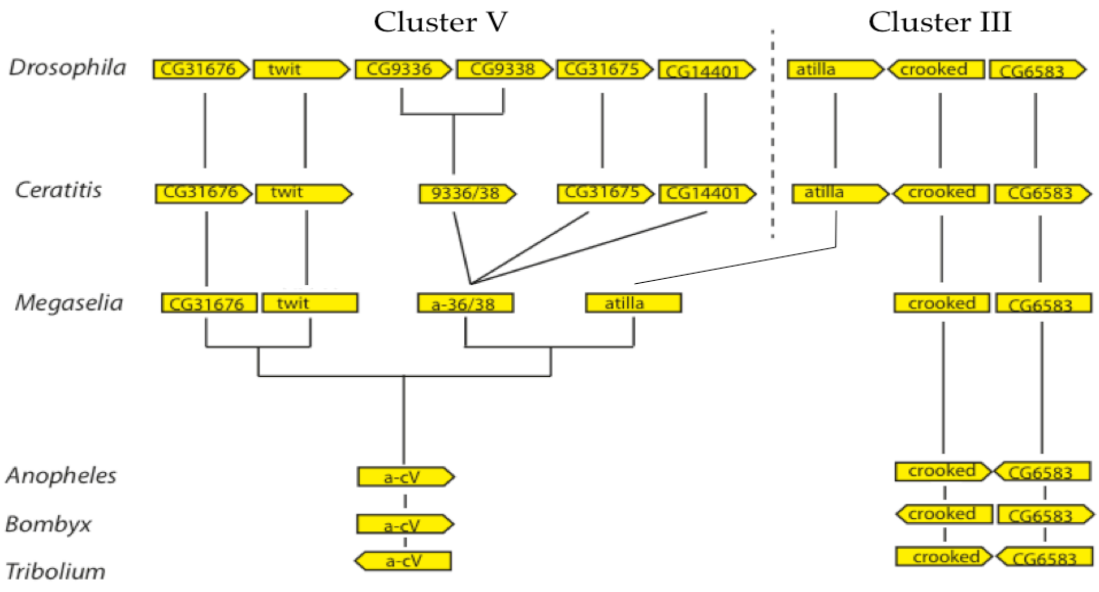

Figure 5. Cluster III and V Ly6 genes evolutionary history. The diagram shows all the ancestral genes known so far and which genes they give rise to at each duplication event. The split into two clusters is depicted with a stitched line.

This study also found that, for each duplication step, one of the duplicates retained the ancestral expression pattern while the other acquires a new one (unpublished data). Gain of novel expression patterns was also observed in copies that inherited the ancestral pattern, resulting in an impressive increase of domains in which these genes are expressed. These results indicate that evolution by neofunctionalization is seemingly behind many of the functional divergence observed in this gene subset of the Ly6 superfamily.

3.1. Evolutionary history and expression pattern of CG9336 and CG9338

CG9336 and *CG9338* are the most recently duplicated paralogues of this model, arising from a single gene of which an orthologue, here called 36/38, can be found in Ceratitis. Observation of the expression patterns of these genes was characterized in our lab (unpublished data). *In situ* hybridization was performed for all three genes, however, *CG9336* and *CG9338* were largely overlapping, due to the probes' high sequence similarity. Thus, YFP trap lines for both genes were used to resolve this problem.

Ceratitis 36/38 expression is more prominent in the Bolwig's organ (BO; a cluster of cells that develops into the larval photoreceptors), trachea, neurons in the midline of the ventral nerve chord (VNC), and glia in the exit nerves (Fig. 7J-7L). This gene also has high expression in epidermis (Fig. 7L), anal plate, pharynx and pharyngeal muscles (Fig. 7J). Expression in the hindgut is also clear (Fig. 7K).

Despite their recent origin, *CG9336* and *CG9338* have very disparate expression patterns. *CG9336* acquired most of Ceratitis 36/38 expression pattern, whereas *CG9338* diverged. *CG9336* is highly expressed in BO, trachea, CNS and peripheral nervous system (PNS). Expression in the CNS is restricted to the midline glia, perineural glia and some dispersed neurons (Fig. 6A). PNS expression is confined to the glial cells in the exit nerves and peripheral sensory organs (Fig. 6A). Expression in the anal plate can also be seen, as well as a novel expression in the posterior spiracles. The epidermal expression seems to be restricted to dorsal stripes in the anterior segments at late stages. Very low levels of expression can be seen in the hindgut. The most striking novelty is the prominent expression in the cardiac tube (cardioblasts and pericardial cells of heart and aorta), and low levels of expression can also be seen in the lymph gland (Fig. 6B).

CG9338 has lower levels of expression in general, when compared with *CG9336*. Low levels of expression can be seen in the same midline and PNS glial cells observed in *CG9336* expression pattern (Fig. 6C). The most prominent expression is detected in the migrating hemocytes, which are cells of the immune system in insects (Fig. 6D). This expression is a striking novelty in this gene lineage.

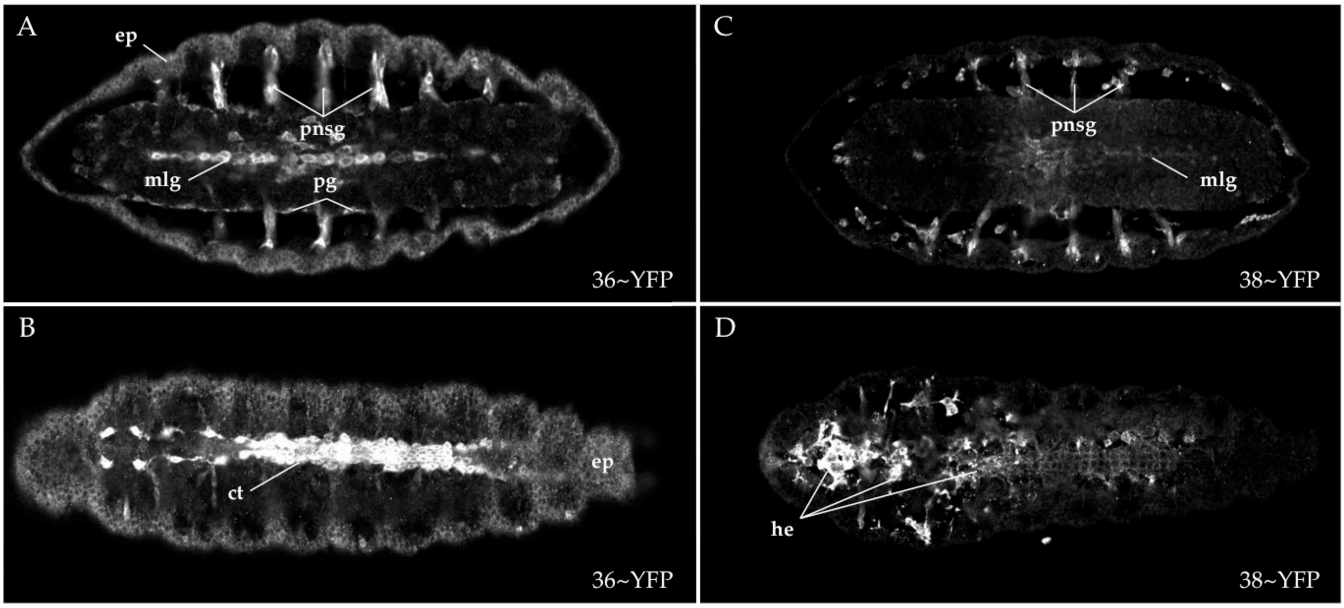

Figure 6. Expression patterns of *D. melanogaster CG9336* and *CG9338*. Expression of YFP fusion proteins is shown. (A, B) 9336~YFP line, expression in PNS (pnsg), perineural (pg) and midline (mg) glia, cardiac tube (ct) and epidermis (ep) is shown. (C, D) 9338~YFP line, expression in PNS (pnsg) and midline (mg) glia, and hemocytes (he) is shown.

Interestingly, *CG9336* and *CG9338* co-orthologues in closely related Drosophilids show expression pattern differences between species. *In situ* hybridization was done in our lab for *D. virilis*, *D. pseudoobscura* and *D. ananassae*. *D. virilis* shows lack of expression in the cardioblasts and very weak expression in the pericardial cells, both highly expressed in *D. melanogaster*, and had a much more prominent and broad epidermal expression, similar to the one seen in *Ceratitis* (Fig. 7G-7I). Midline glia expression is also similar to *Ceratitis*. In *D. pseudoobscura*, expression in epidermis is still observed with lower levels, and some expression could be seen in the most anterior part of the cardiac tube (Fig. 7D-7F). In general, *D pseudoobscura* expression pattern seems to be a mix of *D. melanogaster* and *Ceratitis* expression patterns. *D. ananassae* expression pattern is much more similar to *D. melanogaster*, with visible expression in the cardiac tube, differing only in the epidermal expression, which is restricted to stripes of epidermal cells in all segments, rather than only anterior segments on the dorsal side. *CG9336* and *CG9338* expression patterns are not distinguishable for these species due to the probes' high sequence similarity.

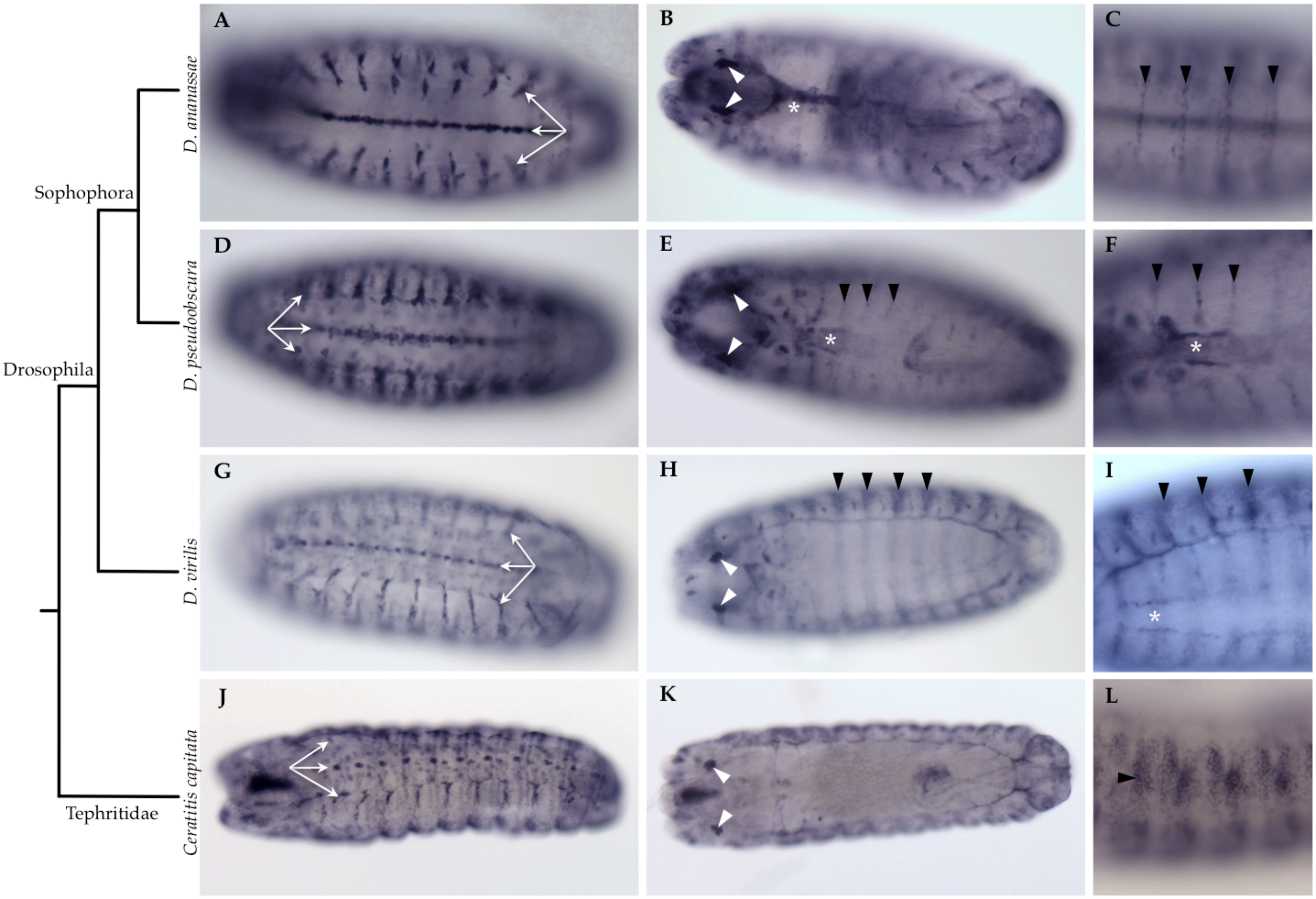

Figure 7. *CG9336* and *CG9338* expression pattern divergence. *In situ* hybridization of duplicates of *D. ananassae* (A-C), *D. pseudoobscura* (D-F) and *D. virilis* (G-I), and of the unduplicated orthologue of *Ceratitis capitata* (J-L). Glia (white arrows), Bolwig's organ (white arrowheads), epidermis (black arrowheads) and cardiac tube (asterisks) expressions are shown.

The evolutionary history of these duplicates and their rapid expression pattern divergence in a small evolutionary time scale, make them a good model to explore *cis*-regulatory evolution and its impact on duplicated genes. In this work, I aimed to locate the enhancers responsible for the divergent expression patterns of *CG9336* and *CG9338*, as a first step to better understand which modifications the *cis*-regulatory elements of these genes underwent to give rise to new expression domains (i.e., functions). I also took advantage of the mVista tool to obtain a sequence conservation map of this region for *D. ananassae*, *D. pseudoobscura* and *D. virilis*, to see if there were any patterns of non-coding sequence conservation that could explain the expression pattern divergence observed between Drosophilids.

II. Materials and Methods

1. *Organism stocks*

All *Drosophila melanogaster* stocks were maintained on standard food at 25ºC. CG9336~YFP (DGRC number 115180) and CG9338~YFP (DGRC number 115071) protein trap lines were obtained from Kyoto Stock Center. attP2 (Stock number 25710) and UAS-mCD8::GFP (Stock number 5137) lines were obtained from Bloomington Drosophila Stock Center.

2. *Sequence analysis and primer design*

CG9336 and *CG9338* 5′/3′ intergenic and intronic sequences were downloaded from modENCODE Genome Browser, by highlighting the region of interest and saving the selection to a FASTA file.

Specific primers for each sequence of interest were designed in Primer3 (v4.0.0) using default settings, except for GC%, which was set from 40 to 60, and analyzed with Netprimer for hairpin, dimer or cross dimer formation.

Sequence alignments of sequencing results were carried out in EBI ClustalW2 using default settings. Sequence conservation map of *CG9336/CG9338* region was obtained by using the mVista tool in the Vista tools website, which designs the map by aligning and comparing sequences from different species. Compared species were: *D. melanogaster, D. ananassae, D. pseudoobscura* and *D. virilis*. Sequences from *D. yakuba, D. erecta* and *D grimshawi* were used to increase the alignment robustness. The sequences were taken from Flybase, each starting in the second exon of CG9335 and ending in the second exon of CG31675 for a better alignment of the sequences.

Fragment name	Forward Primer	Reverse Primer	Length (bp)
14400-9336 IR	TCCGCTTGTTGTCATAGCTG	AGTGTGGGAATTGCTTTTCG	1,459
9336 Int1	TGATCAGTCTGGCTTGTTCG	CTCGCACTGGTAGCACTTGA	441
9336 Int2	GCATGAAGAAGACCCTCGAA	GCAGCCAGCTTGGATATTGT	872
9336-9338 IR1	CGACGAATGGTTGTCATAGC	GTGCCGTAGACGTTCCAGTT	2,900
9336-9338 IR2	GATCCTTTTGATTTGCCAGA	TCATACCAGTGCAGGCTACG	2,974
9338 Int2	CACCTCGATTCCTCCAGAAC	GCCAAAGTAGCAGGACCTCA	1,460
9338-31675 IR	AAACTGTTTTCGCCCACATT	AATTCTTGGAGCCGTCTGTC	539

Table 1. *D. melanogaster* primer sequences and fragment size for each tested region.

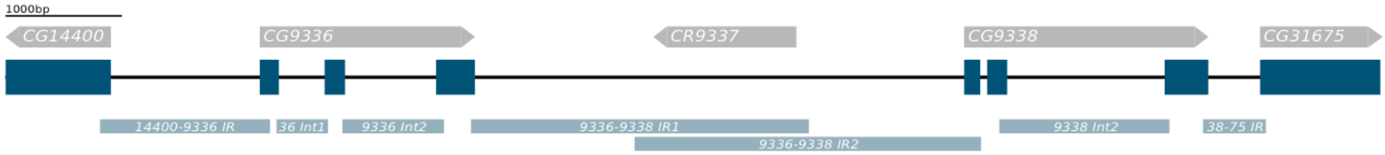

Figure 1. Candidate regions used to test enhancer activity. Each isolated fragment used to create the reporter constructs is mapped and named. Sizes are fit to the scale given. Grey arrows – genes (pseudogene in case of CR9337), Dark blue boxes – exons, Light blue boxes – candidate regions.

3. Fragment isolation and construction of reporter genes

CG9336 and *CG9338* 5'/3' intergenic and intronic sequences were PCR amplified from *D. melanogaster* Oregon R genomic DNA, using the region specific primers designed above (Table 1). The PCR products were then run on a 2% agarose gel to isolate the fragment of interest through gel purification, ligated to pGem-T Easy Vector (Promega) and transformed into *E. coli*. Resulting colonies were screened by PCR, using T7 and SP6 primers (Table S1), and positive clones were sequenced to check for PCR introduced mutations. Clones with minimum errors were chosen and inserted in Gateway pENTR1A Dual Selection Vector (Invitrogen) using the vectors NotI restriction sites.

pENTR specific primers (Table S1) were used to screen for clones with the sequence of interest and orientation was verified by sequencing. Clones with the desired orientation were transferred to the pBPGAL4.2::VP16Uw reporter vector by means of Gateway LR Clonase II (Invitrogen) reaction, and transformed. The Gal4 of this reporter vector has a modification in its activation domain, which augments the levels of expression (Pfeiffer *et al.* 2010*)*. Colonies were then screened by PCR with VP16 specific primers (Table S1) and positive clones were sequenced to confirm sequence identity.

Each construct was purified and injected into embryos of the *w-*; attP2 line of *D. melanogaster*, which has a stable expression of φC31 site-specific integrase controlled by a *nanos* promoter (germ line specific), mediating the integration of the reporter vector into the *attP* site on the 3rd chromosome. An "unmodified" VP16 (i.e., a VP16 plasmid with an intact cassette) construct was also injected as a control for background levels of Gal4 expression. Injected embryos were kept on standard food at 25ºC. Emerged adults were crossed with *yw* flies (i.e., flies mutant for *yellow* and *white* genes, which cause them have a yellower body color and have white eyes, respectively): each male was crossed to 3 *yw* virgin females and groups of 3 females to 2 *yw* males, and the progeny was screened for red-eyed transformants (*white* rescue by *mini-white* gene in construct).

4. Analysis of reporter gene expression patterns

Homozygous stocks were established for each transgenic line by crossing heterozygous F1 flies and crossing homozygous F2 individuals (stronger eye pigmentation). Homozygous males for each line were then crossed with UAS-mCD8::GFP virgin females in cages with apple juice medium plates to collect embryos. Embryos were collected from 3h egg lays and staged at a 17ºC incubator (21-24h: Stage 12; 24-27h: Stage 14; 24-27h + 2h rt: Stage 17). Embryos were dechorionated in 50% bleach, fixed for 30 min in 9.4% fixative (1 ml 200mM EGTA, 1ml 37% formaldehyde, 400 µl 10xPBS, 1.6ml H_2O), devitellinized by methanol shock, washed 4 times with methanol, then 4 times with PBT (0.1% Tween in PBS) and let stand overnight in 200µl of 70% Glycerol in PBS at 4ºC before mounting.

5. Immunohistochemistry

9336~YFP and 9338~YFP embryos were collected and fixed as above. After blocking with 5% normal Goat serum in PBT for 30 min, embryos were incubated with primary antibody overnight at 4ºC. Embryos were then washed 4 times for 20 min in PBT and incubated in the dark with secondary antibody overnight at 4ºC. Embryos were washed again 4 times 20 min in PBT and let stand overnight on 200 µl of 70% Glycerol in 1xPBS at 4ºC before mounting. The primary antibodies used were: Rabbit anti-GFP (1:1000; Molecular Probes, Invitrogen), Mouse anti-Repo (1:50; DSHB), and Mouse anti-Elav (1:500; DSHB). The secondary antibodies used were 488 anti-Rabbit (1:1000; Molecular Probes, Invitrogen) and 546 anti-Mouse (1:1000; Molecular Probes, Invitrogen).

6. Image acquisition

Samples were imaged on an inverted Leica SP5 confocal microscope using a 20x 0.70NA dry objective lens and a 40x 1.25NA oil objective lens. All images were taken with 70% laser power, 1200 gain, a default pinhole size (1 Airy unit), a frame average of 6 and a z-step size of 2 μm for 20x and of 1 μm for 40x. Images were all processed using the Fiji software.

III. Results

1. Characterization of CG9336 and CG9338 expression patterns

To confirm the expression domains of *CG9336* and *CG9338* previously described, a rigorous characterization of the duplicates expression patterns was done. For this purpose, YFP trap lines available for both genes were used. These lines have the YFP-coding sequence inserted in the second introns of each gene, producing YFP-fusion proteins. The YFP expression domains are the same as the ones observed with *in situ*, except for the epidermal expression, which has a wider spatiotemporal domain in 9336~YFP.

As already described before, 9336~YFP was mainly expressed in glia (PNS, midline and perineural), trachea, Bolwig's organ and cardiac tube (Fig. 1A, 1B, 1D, 1E and 1F). High expression levels were also observed in the anal plate and posterior spiracles (Fig. 1C and 1H), and lower levels of expression in epidermis, eye-antennal discs (EAD) and lymph gland (Fig. 1A, 1D, 1F and 1G). I could also identify a group of cells with high expression, located near Bolwig's at younger stages and migrating to the tip of the embryo at final stages, as being the antennal organ (AO) or cells associated with it (Fig. 1E).

For 9338~YFP, expression in hemocytes was visible from stage 14 to stage 17 (Fig. 1I and 1J). As previously observed, low levels of expression could be seen on the PNS glia, as well as in the midline glia (Fig. 1K). Expression on the antennal organ was also visible, but the levels of expression were lower than in 9336~YFP (Fig. 1I and 1L). At stage 17, a dorsal row of unidentified cells in the mid of the head is visible (Fig. 1L).

2. Localization of enhancers regulating the divergent expression domains

To locate the enhancers responsible for the diversified expression patterns of the most recent duplicates, *CG9336* and *CG9338*, reporter gene with each region of interest were constructed. Candidate regions were defined as all non-coding regions (i.e., intergenic regions and introns) of *CG9336* and *CG9338* genomic region. The first intron of *CG9338* was not included due to its small size (52 bp). All sequences were flanked by their neighboring exons, excluding the intergenic sequence between *CG9336* and

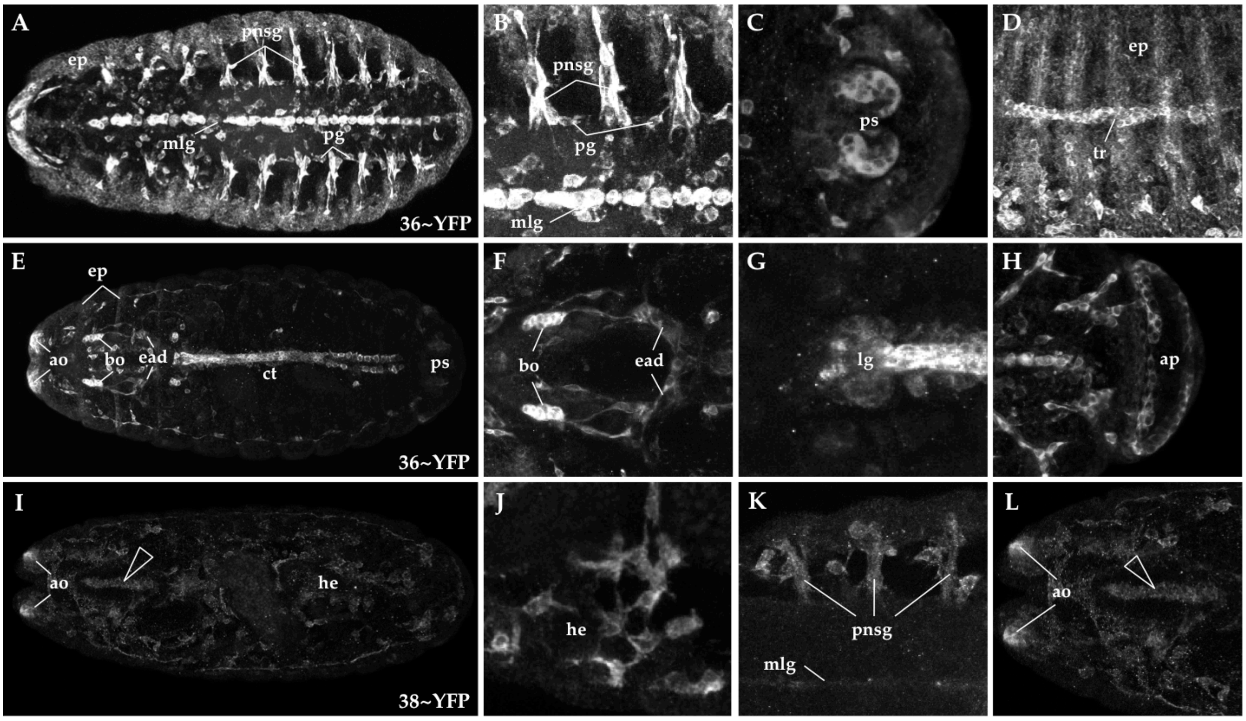

Figure 1 – Expression patterns of *CG9336* and *CG9338*. (A-H) 9336~YFP line (A) Stage 14 embryo, ventral view; (B) PNS glia (pnsg), midline glia (mlg) and perineural glia (pg); (C) Posterior spiracles (ps); (D) Trachea (tr) and epidermis (ep); (E) Stage 17 embryo, dorsal view, antennal organ (ao) and cardiac tube (ct); (F) Bolwig's organ (bo) and eye-antennal discs (ead); (G) Lymph gland (lg); (H) Anal plate (ap). (I-L) 9338~YFP line; (I) Stage 17 embryo, dorsal view; (J) Hemocytes (he); (K) PNS glia and midline glia; (L) Antennal organ and dorsal row of unidentified cells (arrowhead).

CG9338, which was divided into two overlapping sequences with similar size due to the length of the region.

D. melanogaster transgenic lines for each individual reporter gene were generated by means of the φC31 integrase system. These lines were then crossed with a UAS-mCD8::GFP line and the embryos collected and staged to check for embryonic GFP expression patterns by confocal microscopy. This line was chosen because the GFP localizes to the membrane, allowing us to see cell shape. To check for ectopic expression caused by the Gal4 vector and to define basal levels of GFP expression, a transgenic line with an unmodified reporter vector was obtained through the same methods.

All DNA fragments tested for enhancer activity were capable of activating GFP expression above basal levels in at least one of the expression domains of *CG9336* and *CG9338* during the embryonic stages observed. The unmodified Gal4 reporter gene has medium to high levels of expression in subsets of neurons in the CNS and PNS of late stage 17 embryos (Fig. S1B). No expression was observed in all previous stages (Fig. S1A). A summary of the constructs that showed enhancer activity for each of *CG9336* and *CG9338* expression domains follows.

Expression in glial cells (PNS, midline and perineural), cardiac tube, lymph gland, and posterior spiracles were all driven by the 9336-9338 IR1 construct (Fig. 2B, 2E, 2F and 2H). Expression was already visible at stage 12 in PNS and midline glias, and stayed on until stage 17. Perineural glia, cardiac tube and posterior spiracles expressions appeared at late stage 14 and go on until stage 17. Lymph gland expression could only be seen at stage 17. Anal plate expression was only observed in the 9336 Int2 transgenic line at stage 17 (Fig. 2C).

Epidermal expression was only driven by the 14400-9336 IR transgenic line (Fig. 2G). GFP expression could already be seen at stage 12 and remained active until stage 17. The expression was stronger at the ventral side of the embryo and was restrained to rows of epidermal cells. Activation of expression in BO could also be found in the same transgenic line (Fig. 2I). Although, GFP expression was very low or absent in BO before late stage 15. EAD expression was only observed in 14400-9336 IR and 9336-9338 IR2 transgenic lines (Fig. 2I and 2J).

Four distinct constructs activated AO expression: 9336 Int1, 336 Int2, 9336-9338 IR2 and 9338 Int2 (Fig. 2L-2O and 2Q-2S). For all these transgenic lines, expression was already visible in antennal organ at stage 12, on each side of the head. GFP expression pattern migrated in a similar faction as the AO through the stages, ending up at the tip of the embryo at stage 17. Elav immunostaining was performed to check if expression was driven in BO too, as both of these structures are composed of neurons.

Hemocyte expression was found in 9338 Int2 and 9338-31675 IR transgenic lines (Fig. 2U and 2V). Like in the YFP trap line, expression of GFP in hemocytes was visible from stage 14 until stage 17. Also in these transgenic lines, expression in the same dorsal row of unidentified cells of the head, visible in 9338~YFP, was observed (Fig. 2U and 2V).

Transgenic lines capable of activating expression on trachea or hindgut were not found, although, in the 9336 Int2 transgenic line, a similar hindgut expression in the larvae could be seen. A possible explanation for this is that these expressions were masked by others in the same construct, making them difficult to observe.

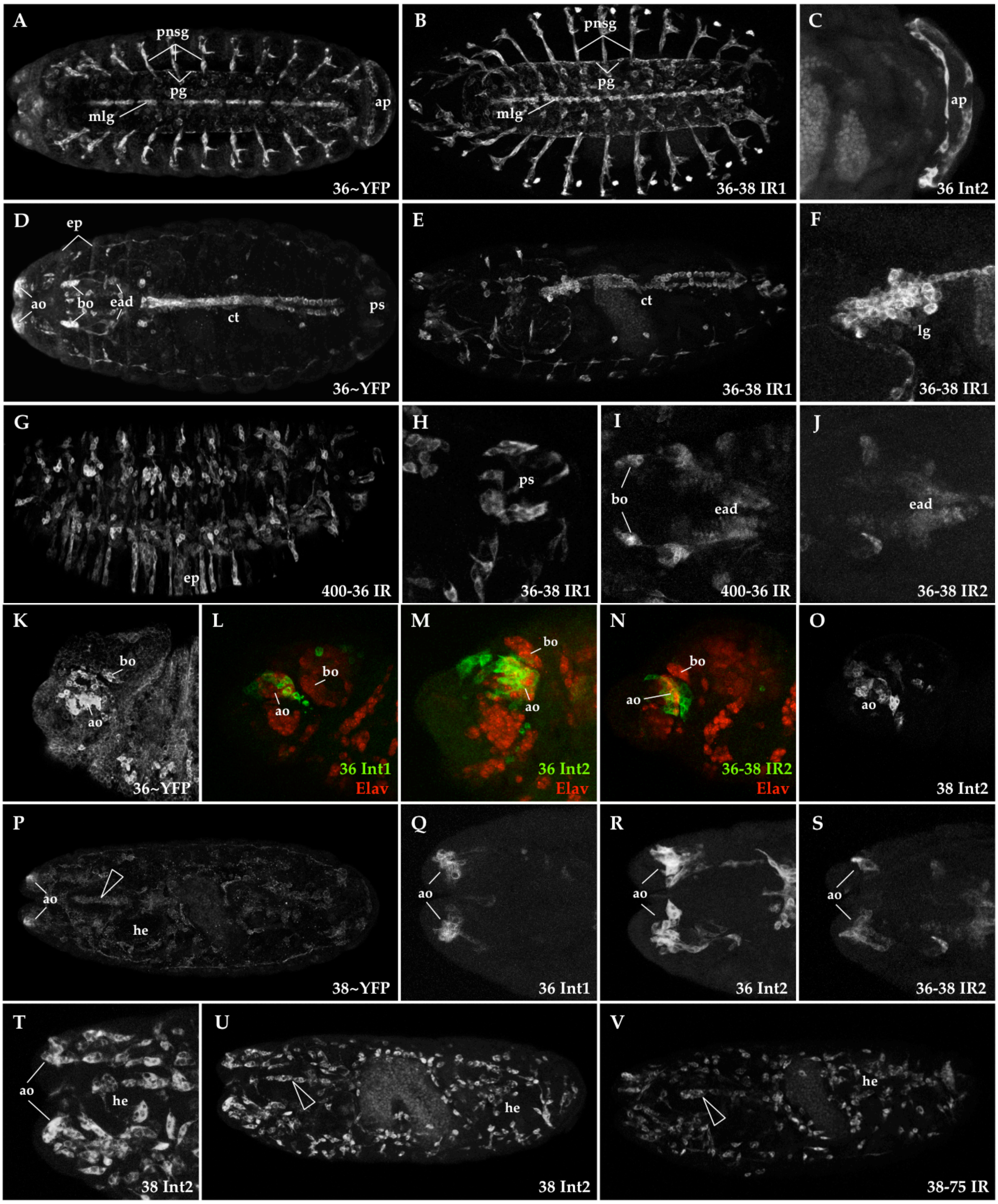

Figure 2 – Reporter constructs expression patterns. (A) 9336~YFP, ventral expression pattern; (B) 36-38 IR1, glial expression; (C) 36 Int2, AP expression; (D) 9336~YFP, dorsal expression pattern; (E, F) 36-38 IR1, cardiac tube and lymph gland expression, respectively; (G) 400-36 IR, epidermal expression; (H) 36-38 IR1, posterior spiracles expression; (I) 400-36 IR, Bolwig's organ and eye-antennal discs expression; (J) 36-38 IR2, eye-antennal discs expression; (K) 9336~YFP, stage 14 antennal organ and Bolwig's organ expression; (L-O) stage 14 antennal organ expression in 36 Int1, 36 Int2, 36-38 IR2 and 38 Int2 lines; (P) 9338~YFP, dorsal expression pattern; (Q-T) stage 17 antennal organ expression in 36 Int1, 36 Int2, 36-38 IR2 and 38 Int2 lines; (U,V) Hemocyte and dorsal line of cells expression in 38 Int2 and 38-75 IR lines, respectively.

3. CG9336/38 expression pattern divergence between Drosophilids and sequence conservation map of 9336/38 genomic region

The expression patterns of the duplicates have undergone some changes throughout the Drosophilids lineage, mainly the epidermal and the cardiac tube expressions. The epidermal expression seems to be progressively lost while cardiac tube expression seems to be progressively gained. Knowing the regions that regulate expression in these tissues in *D. melanogaster*, we can ask if modifications in these regions underlie the changes in expression.

Patterns of sequence conservation can be used to locate enhancers, due to the fact that enhancers tend to be more conserved than non-functional sequences. So, I used the mVista tool to obtain the sequence conservation map of the 36/38 genomic region for the three Drosophilid species for which we know the expression patterns, using *D. melanogaster* as reference: *D. ananassae*, *D. pseudoobscura* and *D. virilis*. The objective behind this was to 1) seek if there was any correlation between the divergent expression patterns observed between drosophilids, and differences in conservation peaks, taking into account the results from reporter genes, and 2) possibly give us some idea of which modifications the enhancers located in the tested regions could have undergone.

In the 5′ intergenic region of *CG9336*, a single peak of about 70-75% conservation is found for *D. ananassae* as well as two peaks, one in the same position as the latter, of 55-60% conservation for *D. pseudoobscura*. There are no peaks of sequence conservation above 50% for *D. virilis* in this region. No conservation peaks of interest were observed in the first intron of *CG9336*. It’s worth noting that this intron is much smaller in *D. pseudoobscura* and *D. virilis*, making it probable that it does not contain any enhancer. For the second intron of *CG9336*, significant conservation peaks were only detected in *D. pseudoobscura*.

The huge gap of sequence conservation observed in the intergenic region between *CG9336* and *CG9338* is caused by *CR9337*, a pseudogene only present in *D. melanogaster*. Close to *CG9336*, two high conservation peaks can be identified for *D. ananassae* and *D. pseudoobscura*. In *D. virilis* only the left peak, with not much more than 50% conservation, can be seen. At 5′ of *CG9338*, small conservation peaks, with 50-60% sequence conservation, can be found for *D. ananassae, D. pseudoobscura* and *D. virilis*. The second intron of

CG9338 shows peaks of sequence conservation in *D. ananassae* and *D. pseudoobscura*, but no peaks are found for *D. virilis*. Lastly, in the 3′ intergenic region of *CG9338*, two conserved peaks are visible in *D. ananassae* and *D. pseudoobscura*, although, in *D. virilis* only the rightmost one is still noticeable.

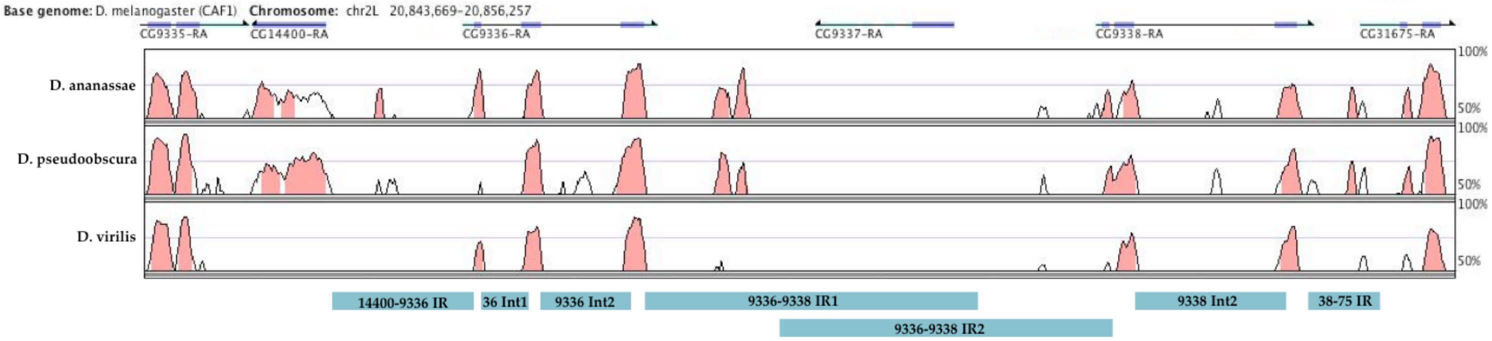

Figure 3 – Sequence conservation maps of 9336/38 region. Resulting conservation maps of the sequence conservation analysis, performed in mVista, of the genomic region of *CG9336* and *CG9338* for *D. ananassae*, *D. pseudoobscura* and *D. virilis*. Name and relative positions of each construct is given at the bottom.

IV. Discussion

1. Cis-regulatory architecture of CG9336 and CG9338

All the tested non-coding regions were able to drive expression in at least one of the duplicates expression domains, leading to the general conclusion that enhancers regulating these genes can be found in all the non-coding sequences of the 9336/38 genomic region, having a scattered genomic location in relation to the gene they regulate.

Enhancers controlling epidermal, BO and EAD expression can be found in the 5′ Intergenic region of *CG9336*. The first intron of *CG9336* only contains one enhancer controlling antennal organ expression. The second intron of *CG9336* has enhancers controlling AO and anal plate expressions. The 3′ intergenic region of *CG9336* contains enhancers controlling expression in glial cells, cardiac tube, lymph gland and posterior spiracles. The intergenic region closer to *CG9338* contains enhancers controlling AO and EAD expressions. Both the second intron and the 3′ Intergenic region of *CG9338* contain enhancers regulating expression in the hemocytes and the dorsal line of cells in the head. An enhancer regulating AO expression can also be found in the second intron of *CG9338*.

2. Evolutionary history of the cis-regulatory elements regulating conserved expression domains

From *C. capitata* to *D. melanogaster*, all five species show conservation of the Bolwig's organ and glial cells expression patterns. Such conservation is expected to be observed in the enhancers regulating these expressions.

In the 5′ region of *CG9336*, which drives Bolwig's organ expression, an absence of conservation peaks observed for *D. virilis* contradicts the idea of a conserved enhancer, even more so if we assume that the peaks observed in *D. ananassae* and *D. pseudoobscura* correspond to the enhancer that regulates epidermal expression. It has already been shown that enhancers controlling conserved expression domains can diverge in their genomic locations in relation to the gene they regulate. This is the case of enhancers regulating the *yellow* gene in close related Drosophilids [26]. So, the enhancer controlling BO expression in other species can possibly be in a different genomic location. The enhancer controlling anal plate expression seems to follow a similar trend, as only low conservation peaks are observed for *D. pseudoobscura* in the second intron of *CG9336*.

Glial cells expression is driven by the 3′ intergenic region of *CG9336*. In this region, peaks of high sequence conservation show up for *D. ananassae* and *D. pseudoobscura*, but for *D. virilis*, only a small conservation peak can be seen. A possible scenario is that the position of the enhancer itself is conserved but the enhancer went through some modifications after the split of Sophophora from the rest of the Drosophilds. This could explain the small difference observed in the shape of the VNC midline cells between *D. virilis* and *D. pseudoobscura*. This enhancer regulates prominently *CG9336* expression, but it seems that it is close enough to *CG9338* to still drive some expression in glia. There's a possibility that this expression was driven in both duplicates in other species, but due to the insertion of the *CR9337* pseudogene, the distance of the enhancer to *CG9338* was disturbed, possibly causing it to regulate primarily *CG9336*. The glial enhancer was possibly co-opted to drive the posterior spiracles expression, as these cells can be glia associated to sensory organs. In the case of the EAD expression, it is possible that an enhancer 5′ of the unduplicated orthologue driving this expression was duplicated along with the gene, as this expression is observed in both 5′ regions of each duplicate in *D. melanogaster*.

3. Evolutionary history of the cis-regulatory elements regulating divergent expression domains

Epidermal expression has continuously changed through the duplicates evolutionary history, going from being highly expressed in *Ceratitis* to being expressed at low levels in anterior segments of *D. melanogaster*. This could be explained by the constant modification of the enhancer regulating this expression. The conservation peaks observed in the region driving expression in epidermis support this explanation, since they seem to lose conservation along with the change in expression (i.e., the enhancer possibly undergone modifications). Although, reporter epidermal expression was not similar to the one seen with *in situ* hybridization, meaning that the enhancer contained in this region seems to be insufficient to regulate expression in that restricted domain. The final expression pattern can be the result of post-transcriptional regulation (e.g., mRNA degradation), or the interaction of a repressor with the enhancer. Changes in any of these mechanisms can modify the expression pattern observed.

The 3′ intergenic region of *CG9336* regulates cardiac tube expression, which, similarly to the epidermal expression, went through continuous changes in the Drosophilid lineage. *De novo* formation of an enhancer driving low levels of expression in pericardial cells, during the divergence time between Ceratitis and *D. virilis*, and continuous co-option of this enhancer for cardiac tube expression through the Drosophilid lineage, could explain this. The right peak of conservation in this region seems to follow this trend, as it has higher conservation in *D. ananassae*, a bit lower conservation in *D. pseudoobscura* and no peak in *D. virilis*, depicting continuous modifications through the evolution of *CG9336*. Lymph gland expression is probably the result of co-option of this enhancer, since this structure is closely related to the pericardial cells [27].

Hemocyte expression is a striking novel expression found in *CG9338*. Two enhancers in different regions drive this expression: one in the second intron and another in the 3′ intergenic region of *CG9338*. Presence of peaks of conservation in both regions for *D. ananassae* and *D. pseudoobscura*, but absence of peaks in the intron and loss of the higher conserved peak in the 3′ intergenic region of *D. virilis*, point out to a scenario where Hemocyte expression arose after the split of Sophophora, by the appearance of these two enhancers. These enhancers were more probably originated by co-option of duplicated enhancers, but *de novo* creation is also a possibility.

4. Evolution of seemingly redundant cis-regulatory elements

AO expression seems to be regulated by four enhancers in four different regions. These enhancers drive highly overlapping expression patterns, making them apparently redundant. However, recent studies have shown that enhancers with overlapping activities can contribute to phenotypic robustness, being maintained by natural selection [28]. The enhancers in the first and second introns of *CG9336* probably regulate expression of *CG9336* in this organ, while the enhancers in the 5′ intergenic region and second Intron of *CG9338* control expression of *CG9338*. Presence or absence of this expression in other Drosophilids and Ceratitis has yet to be checked.

The enhancer in the first intron of *CG9336* is probably related to the enlargement event of this intron, which took place somewhere between *D. pseudoobscura* and *D. ananassae*. This could be the case of an insertion of a transposable element containing an enhancer, which was co-opted to drive AO expression. Some degree of conservation in the second intron of *CG9336* is only observed in *D. pseudoobscura*, making it less probable that the enhancer controlling AO expression is conserved. Although, the presence of an enhancer controlling the same expression in the second intron of *CG9338*, makes it possible that an enhancer that drove expression in the AO was duplicated along with the gene and maintained in both duplicates. Similarly, an ancestral enhancer 5′ of the unduplicated orthologue could be duplicated and co-opted, giving rise to the enhancer 5′ of *CG9338*.

VI. Conclusion and Future directions

Cis-regulatory evolution seems to have had a sizeable role in the evolution and divergence of the expression patterns of *CG9336* and *CG9338*, although, the extent of this role is far from understood, as the evolution of the *cis*-regulatory modules regulating these genes is not as linear as it was thought to be. These genes seem to have undergone many *cis*-regulatory element modifications just within the Drosophilids, and probably even more between *Ceratitis* and *D. melanogaster*. There seems to be a trend of duplicated enhancers that are co-opted to drive expression in new expression domains, which may have made possible the maintenance of these duplicates.

The data from this study alone is not sufficient to discern which modifications in the *cis*-regulatory elements regulating these genes underlie the changes in expression observed between species, as they only show us a single state of their *cis*-regulatory evolution. For that, a rigorous analysis of the *cis*-regulatory modules controlling the duplicates' expression patterns at different evolutionary time points is needed, as well as the *cis*-regulatory modules present in the unduplicated gene state found in Ceratitis. Also, the fact that these genes have no known function makes it difficult to distinguish which of the expression domains are of greater importance (i.e., affect function if changed). A functional analysis would help addressing the question of which *cis*-regulatory changes could be behind the maintenance and fixation of the duplicates.

A similar approach to the one used in this work can be used to analyze the *cis*-regulatory modules regulating these genes in the species *D. ananassae*, *D. pseudoobscura* and *D. virilis*. This would require the isolation of the orthologous candidate regions for each of these three species, construction of reporter vectors with each of these fragments and transformation of these constructs into *D. melanogaster* to check for enhancer activity. By incorporating these vectors in a *D. melanogaster* genetic background, it would be possible to discern *cis*- from *trans*- evolution. These results could be then compared with the results obtained in this work to check if there were modifications in the *cis*-regulatory background that could explain the changes observed between these species. In fact, I have initiated the cloning of the regions driving expression in Bolwig's organ, epidermis, glial cells, cardiac tube and hemocytes for *D. ananassae, D, pseudoobscura* and *D. virilis* (i.e., 5' and

3′ intergenic regions of *CG9336*, and second intron and 3′ intergenic region of *CG9338*). The reasoning behind the choice of the regions is that they drive expression in the most strikingly conserved (Bolwig's organ and glia) and divergent (epidermis, cardiac tube and hemocytes) expression domains. These clones will be used in the follow-up studies to elucidate how these *cis*-regulatory modules have evolved to give rise to the expression patterns we find in each co-orthologue. This analysis could possibly be extended to the Ceratitis unduplicated orthologue, adding information on the ancestral state of the *cis*-regulatory modules.

VII. References

1. Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20: 1313-1326.

2. Ohno S (1970) Evolution by gene duplication. Sringer Verlag, Berlin.

3. Hurles M (2004) Gene Duplication: The Genomic Trade in Spare Parts. PLoS Biol 2(7): e206 doi:10.1371/journal.pbio.0020206.

4. Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11: 97-108.

5. Zhang J, Zhang YP, Rosenberg HF (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet 30: 411-415.

6. Zhang J (2003) Evolution by gene duplication. Trends Ecol Evol 18: 292-298.

7. Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: Glimpses from the young and old. Nat Rev Genet 4: 865-875.

8. Conant GC, Wolfe KH (2008) Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet 9: 938-950.

9. Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlthwait J (1999) Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics 151: 1531-1545.

10. Lynch M, Force A (2000) The Probability of Duplicate Gene Preservation by Subfunctionalization. Genetics 154: 459-473.

11. Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc Biol Sci Lond B 256: 119-124.

12. Wittkopp P, Kalay G (2012) *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet 13: 59-69.

13. Frankel N, Erezyilmaz DF, McGregor AP, Wang S, Payre F, Stern DL (2010) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature 474: 598-603.

14. Shirangi TR, Dufour HD, Williams TM, Carroll SB (2009) Rapid Evolution of Sex Pheromone-Producing Enzyme Expression in *Drosophila*. PLoS Biol 7(8): e1000168 doi: 10.1371/journal.pbio.1000168.

15. Williams MT, Selegue JE, Werner T, Gompel N, Kopp A, Carrol SB (2008) The regulation and evolution of a genetic switch controlling sexually dimorphic traits in *Drosophila*. Cell 134(4): 610-623.

16. Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, Carrol SB (2008) The evolution of gene regulation underlies a morphological difference between two *Drosophila* sister species. Cell 132(5): 783-793.

17. Wray GA (2007) The evolutionary significance of *cis*-regulatory mutations. Nat Rev Genet 8: 206-216.

18. Cande J, Goltsev Y, Levine MS (2009) Conservation of enhancer location in divergent insects. Proc Natl Acad Sci USA 106: 14414-14419.

19. Eichenlaub MP, Ettwiller L (2011) De novo genesis of enhancers in vertebrates. PLoS Biol 9(11): e1001188. doi: 10.1371/journal.pbio.1001188.

20. Rebeiz M, Jikomes N, Kassner VA, Carrol SB (2011) Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. Proc Natl Acad Sci USA 108: 10036-10043.

21. Cande JD, Chopra VS, Levine M (2009) Evolving enhancer-promoter interactions within the tinman complex of the flour beetle, *Tribolium castaneum*. Development 136: 3153-3160.

22. Galat A, Gross G, Drevet P, Sato A, Ménez A (2008) Conserved structural determinants in three-fingered protein domains. FEBS Journal 275: 3207-3225.

23. Hijazi A, Masson W, Augé B, Waltzer L, Haenlin M, Roch F (2009) *boudin* is required for septate junction organization in *Drosophila* and codes for a diffusible protein of the Ly6 superfamily. Development 136: 2199-2209.

24. Nilton A, Oshima K, Zare F, Byri S, Nannmark U, Nyberg KG, Fehon RG, Uv AE (2010) Crooked, Coiled and Crimpled are three Ly6-like proteins required for proper localization of septate junction components. Development 137: 2427-2437.

25. Kim NC, Marqués G (2012) The Ly6 neurotoxin-like molecule target of wit regulates spontaneous neurotransmitter release at the developing neuromuscular junction in *Drosophila*. Devel Neurobio 72: 1541-1558.

26. Kalay W, Wittkopp PJ (2010) Nomadic enhancers: tissue-specific *cis*-regulatory elements of *yellow* have divergent genomic positions among *Drosophila* species. PLoS Genet 6(11): e1001222. doi: 10.1371/journal.pgen.1001222.

27. Ponzielli RP, Astier M, Chartier A, Gallet A, Thérond P, Sémériva M (2002) Heart tube patterning in *Drosophila* requires integration of axial and segmental information provided by the *Bithorax Complex* genes and *hedgehog* signaling. Development 129: 4509-4521.

28. Fujioka M, Jaynes JB (2012) Regulation of a duplicated locus: Drosophila *sloppy paired* is replete with functionally overlapping enhancers. Dev Biol 362: 309-319.

Supplemental data

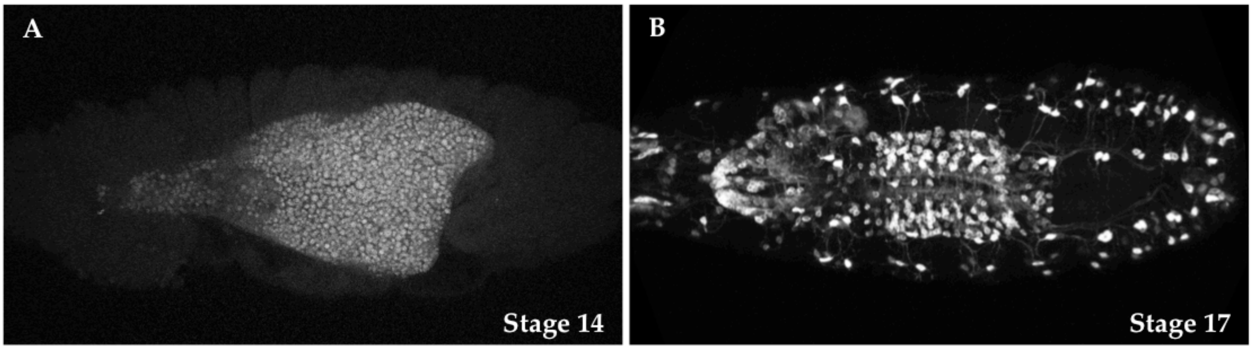

Figure S1. Unmodified Gal4 vector activity. (A) Early stage embryo, no expression visible at stages lower than 17. (B) Late stage 17 embryo, expression in the nervous system.

Vector	Forward Primer	Reverse Primer
pGem	TAATACGACTCACTATAGGG	ATTTAGGTGACACTATAG
pENTR	TTAGTTAGTTACTTAAGCTCGGG	GTAACATCAGAGATTTTGAGACAC
VP16	CCATTATTATCATGACATTAACC	CGTTTATCACCACTTTGTACAAG

Table S1. pGem, pENTR and VP16 vector primers.