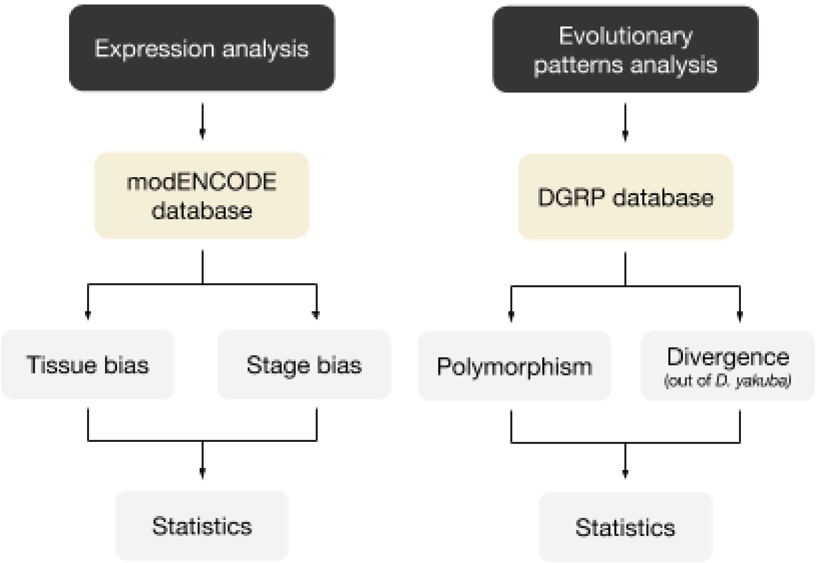
In the last report (24/05/2017) we retrieved all the possible paralogs pairs of *Drosophila melanogaster* from the OMA database: a total of 460 paralog pairs that have one domain missing either in the back or in the front of the protein and 14,005 paralog pairs that share the same domain rearrangement as a control group. During the last two weeks, we have performed two different preliminar analysis: **1)** a **differential expression analysis** and **2)** an **evolutionary pattern analysis**.



1. Differential expression analysis

Expression data of 18,157 genes was downloaded from FlyBase (release Dmel_R6.15 April 2017; file: "gene_rpkm_report_fb_2017_02.tsv.gz"). The downloaded file reports gene expression values as reads per kilobase per million reads (RPKM). RPKM values are calculated only for the unique exonic regions of the gene (excluding segments that overlap with other genes), except for genes derived from dicistronic/polycistronic transcripts, where in which case all exon regions were used for the estimation of RPKM expression calculation. RPKM are calculated using the method from Motazavi et al. 2008.

The dataset consists of expression data for 30 stages of the whole life cycle of *D. melanogaster*, including 12 embryonic samples collected at 2-h intervals for 24 h, six larval, six pupal and three sexed adult stages at 1, 5 and 30 days after eclosion. Regarding the tissue expression data, it consist on the following tissues sampled at different stages of the development:

Head	Ovary	Acc. gland	Testis	Carcass	Dig. syst.	CNS	Saliv.	Fat	Imag. disc
F 1d F 4d F 20d M 1d M 4d M 20d VirF 1d VirF 4d VirF 20d	VirF 4d F 4d	M 4d	M 4d	A 1d A 20d A 4d L3 Wand	A 1d A 20d A 20d L3 Wand	P8 L3	WPP	PP P8 L3 Wand	L3 Wand

$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}$$

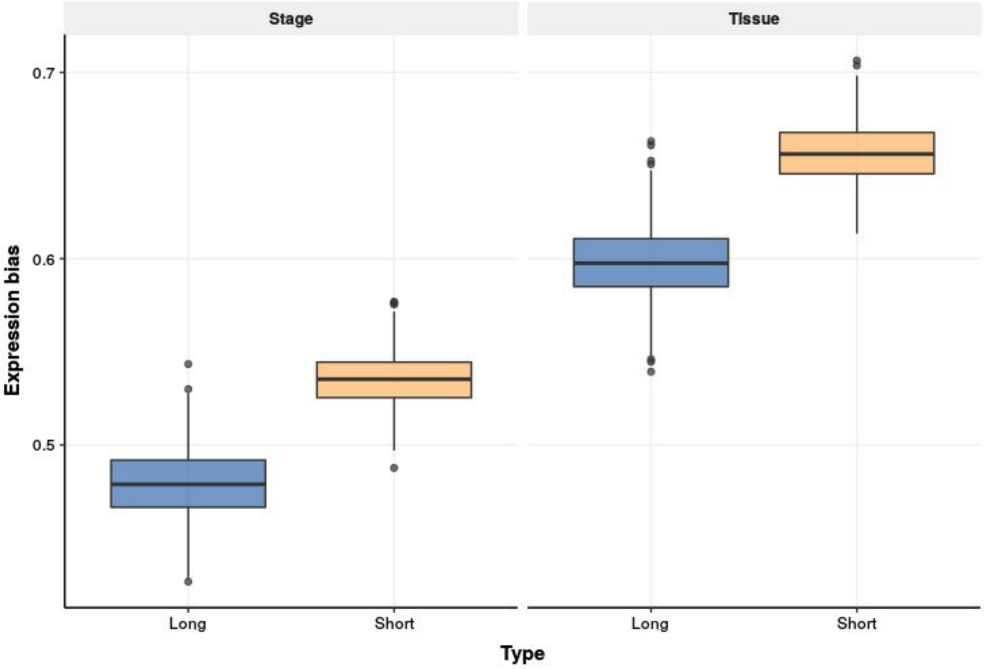
Using the RPKM values provided, we estimate the expression bias for the 30 stages and 29 tissues. We first log-transform the RPKM values (as $log(RPKM + 1)$) and then apply the formula (as Yanai et al. 2005, Larracuente et al. 2008).

In the formula, N is the number of tissues and x_i is the expression profile component normalized by the maximal component value. We will use the FPKM values as a measure of expression. τ ranges from 0 to 1, with values close to 0 indicating broadly expressed genes (housekeeping genes) and values close to 1 indicating genes with a highly biased (or specificity) expression. For example, a gene with a τ = 1, means that is only detectable in one sample (tissue or stage) while τ = 0, that is expressed in all samples with the same expression level.

Expression bias analysis

Are there differences between genes with the short and long domain conformation?

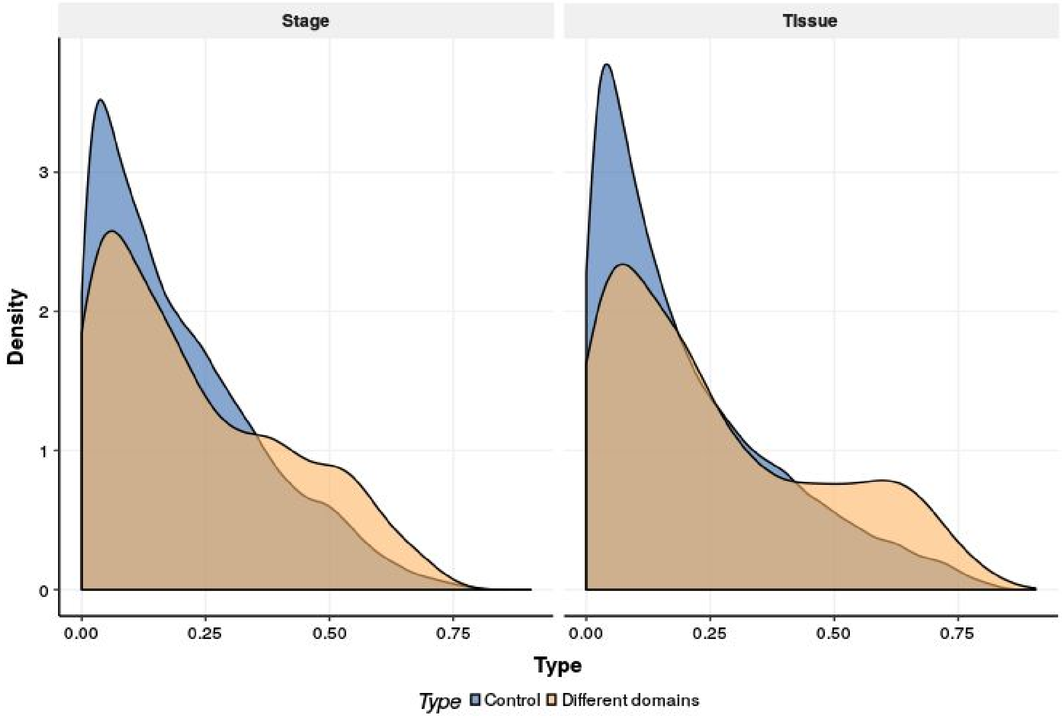
From our 460 paralog pairs with different domains, we first analyse if there are statistical differences between the genes that have the long domain conformation and the genes that have the short domain conformation. In the first group we have 183 unique genes and in the second 256 unique genes. We apply a *t-test* and we find that the genes with the short domain conformation have a higher stage and tissue expression bias than the genes with the long domain conformation (*p* = 0.039 and *p* = 0.016, respectively). The following graph represents sampling distribution of the mean for the two group of genes (bootstrap sample size is 1,000, sampling with replacement):



Comparison with the control group

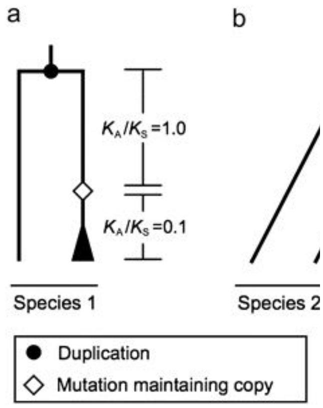
To compare this result with the control group, we calculate the absolute difference of the stage and tissue expression bias between paralog pairs for both pairs with domain rearrangement and control group.

We find that both stage and tissue expression bias absolute difference is higher in the paralog pairs that have a domain rearrangement compared with the control group (Wilcoxon test, p = 0.007 and p = 8.24×10^{7}, respectively). The following graph represents the absolute difference distribution of the two groups.



Because the expression bias absolute difference between paralog pairs with a domain rearrangement is higher in both tissues and stages compared to the control, it indicates that there is a significant change in the expression specificity of paralog pairs with a domain rearrangement. Also, note that the difference is higher for the tissue specificity.

2. Evolutionary patterns

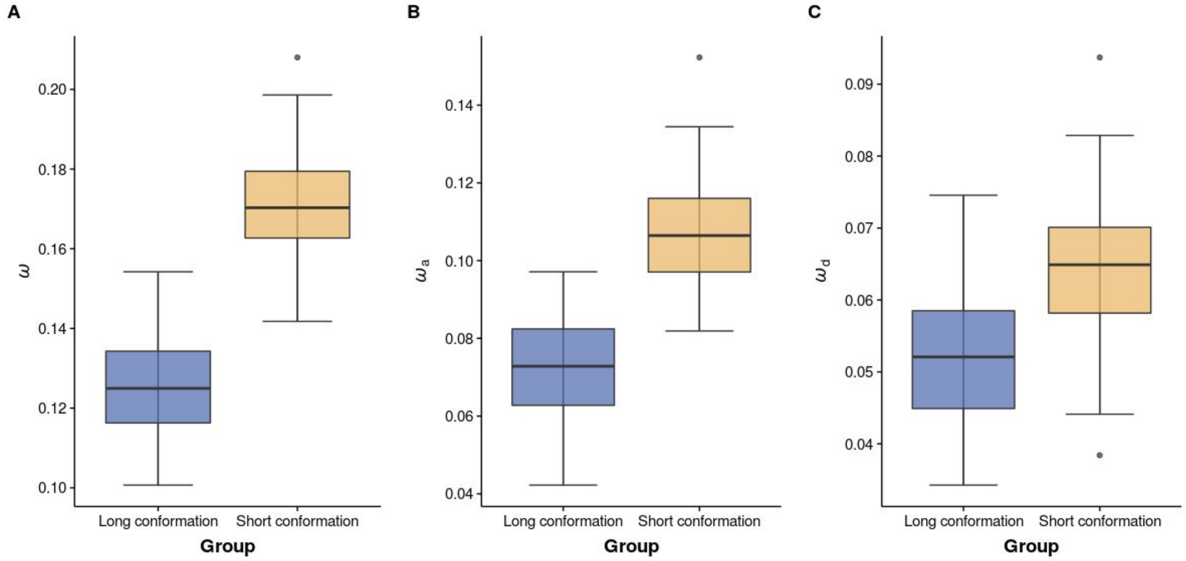


Gene duplications are a source of new genes and protein functions. The innovative role of duplication events makes families of paralogous genes an interesting target for studies in evolutionary biology. From an evolutionary point of view, immediately after gene duplication, the pressure of negative selection weakens during a period of time. One of the proteins encoded by two closest paralogs accumulate substitutions significantly faster than its partner. The McDonald-Kreitman test (MKT) can be positively misleading when applied to the study of duplicated genes. In this graph, it is represented an scenario where 2 paralogs from the same genome are compared. If there is little constraint early in the history of a duplicate (a high K_a/K_s, then many nonsynonymous fixed differences will accumulate because they are neutral. If selection then gets stronger in the recent past, as the duplicated gene can acquire new functions, the ratio of nonsynonymous/synonymous mutations (K_a/K_s) is lower in the present. The excess of fixed nonsynonymous differences can lead to the rejection of the neutral hypothesis, interpreting positive selection in the history of the duplicate gene. But positive selection on the coding sequence has not necessarily occurred as an environmental change. Even the change responsible for the new function may not ocurred in the coding sequence (e.g., a regulatory change).

Evolutionary patterns analysis

For the analysis of the evolutionary patterns, we are going to use DFE-alpha. The DFE-alpha method, is an extension of the classic McDonald and Kreitman test. However, MKT does not take into account the segregation of slightly deleterious alleles, biasing downward the estimation of the adaptive rate. DFE-alpha method corrects for it, providing a more accurate estimation than the MKT and other methods that do not take polymorphism data into account. Briefly, this software uses a maximum-likelihood (ML) method based on polymorphism data to infer the distribution of fitness effects (DFE) of new mutations. It assumes to classes of sites in the genome: neutral sites (synonymous) and selected sites (nonsynonymous) and contrasts the site frequency spectrum (SFS) at these two classes. As a neutral reference, we used the 4-fold degenerate sites and for our target sequence, we used 0-fold degenerate sites. Provided the SFS at both neutral and selected sites together with divergence data, DFE-alpha allows DFE-alpha the calculation of the proportion of fixed adaptive substitutions (α) and the rate of adaptive substitutions relative to the neutral rate (ω_α). Furthermore, in our analysis we include another parameter, ω_d, which represent the proportion of non-adaptive substitutions (slightly deleterious and neutral) relative to the neutral rate.

Of the 460 paralog pairs with domain rearrangements, 183 unique genes maintain the long conformation and 256 the short one. 160 genes out of the 183 and 222 out of the 256 can be analyzed with polymorphism and divergence data using the DGRP database. To estimate these selection parameters, it is necessary to concatenate data from several genes because estimates from a single gene cannot be obtained due to the lack of segregating (or divergent) sites for some site classes. For that, we took 100 samples with replacement of the 160 genes from the group of the long conformation, and 100 samples with replacement of the 222 genes of the group of the short conformation. The following graph represents the sampling distribution of the parameters:



As we can see, the group of genes of the short conformation have a higher ω and ω_α.The CI at 95% doesn't overlap for ω, but it does for ω_α (although the overlapping part is very small). The difference in ω_d is rather subtle.

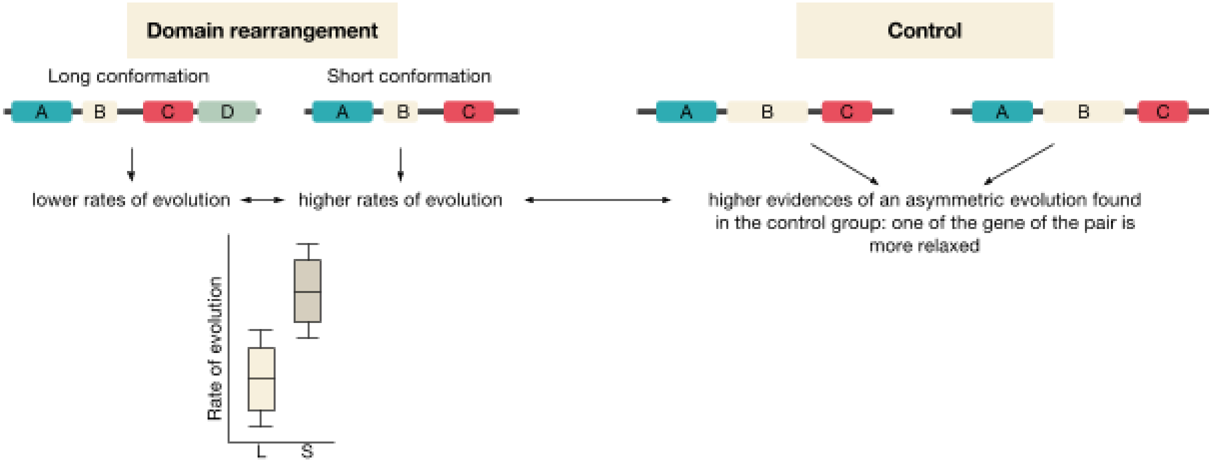
Comparison with the control group

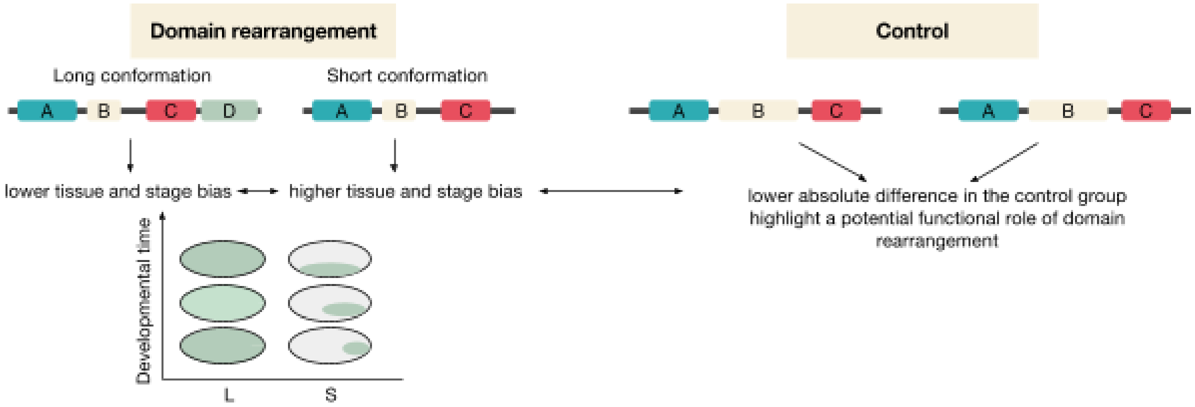
We are going to compare it with our control group. We compute the K_a/K_s for each gene and estimate the absolute difference between the paralog pairs. It is possible to compute it for 353 pairs out of the 460 paralog pairs with domain rearrangements and 9,368 out of the 14,005 of the control group (because there are genes without segregating sites). We find that in the control dataset the absolute difference of K_a/K_s is higher than the pairs with a domain rearrangement (Wilcoxon test, $p = 0.001$).

Summary/Discussion

We have analyzed the expression and evolutionary patterns of paralog pairs that differ in their domain arrangement. Consistent with the idea that one of the proteins encoded by two closest paralogs accumulate amino acid substitutions significantly faster than its partner, we find differences in the rate of evolution between paralog pairs. However, the difference between this evolutionary rate is higher for the control group, suggesting that one of the genes of the pair is under a weaker selective pressure as they have probably not gained a new function. On the other hand, in the group of paralog pairs with domain rearrangements, because one of the genes have lost (or gained) a domain, suggesting the gain of a new function, we expect the difference of evolutionary rates to become smaller, as the gene has became again a target of selection.

For the expression patterns, we also find differences between the two groups. In this case, we find that the genes with a domain rearrangement have a higher difference in expression than the control group, suggesting a functional role of domain rearrangement. It has already been reported a rapid divergence of gene expression profiles for paralog pairs (Chung 2006), but the expression of young paralog has been reported to be highly conserved (Assis 2015).


Domain rearrangement
Control
Long conformation
Short conformation
A
B
C
D
lower rates of evolution
higher rates of evolution
higher evidences of an asymmetric evolution found in the control group: one of the gene of the pair is more relaxed
Rate of evolution
L
S


Domain rearrangement
Control
Long conformation
Short conformation
A
B
C
D
lower tissue and stage bias
higher tissue and stage bias
lower absolute difference in the control group highlight a potential functional role of domain rearrangement
Developmental time
L
S