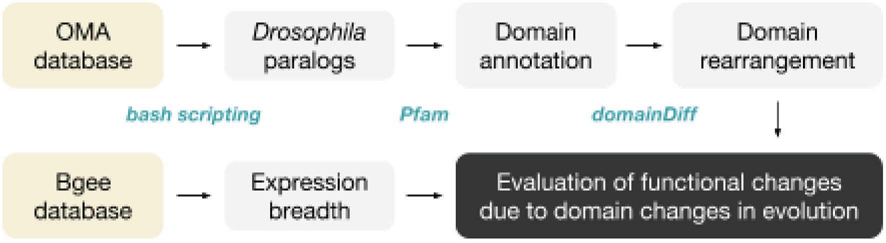
In the first report (12/05/2017) it was settled the first draft of the research plan for the study of the effects of domain modification in *Drosophila*. The research plan consisted on four different parts: **1) orthology and paralogy relationships** from OMA database, **2) domain annotation** with Pfam, **3) domain rearrangement** with **`domainDiff`** and **4) differential expression analysis** from Bgee database.



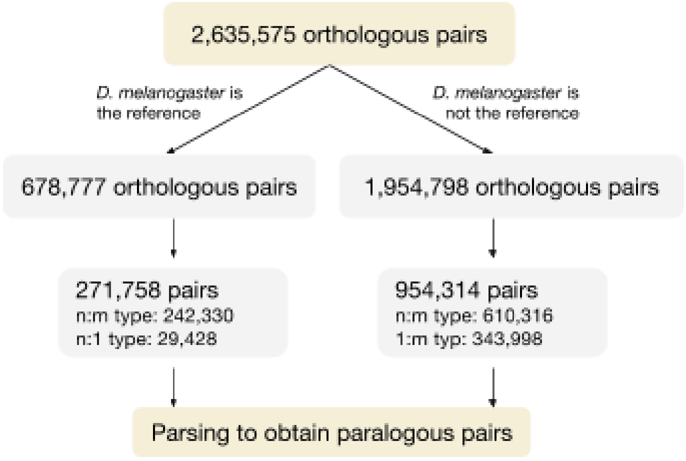
In this report is going to be explained the results from steps 1 to 3, that is to say, from retrieving all the possible paralog pairs from OMA database to obtaining a final candidate paralog pair list with their domain rearrangements to further evaluate their functional impact.

1. Retrieving paralog genes from OMA database

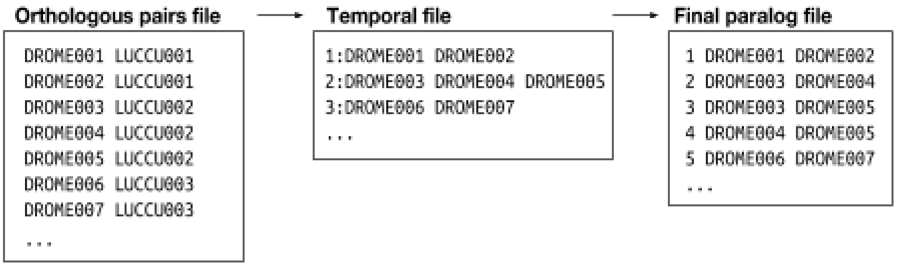
From the current version of OMA database (March 2017) it has been retrieved all the possible *Drosophila melanogaster* orthologous pairs. Specifically, we have obtained a total of **2,633,575 orthologous** to **1,929 different species** from the **`oma-pairs.txt`** file. The list consist on three columns, the two first columns are the orthologous pair genes and the third one indicate the type of orthology (1:1, 1:m, n:1, n:m). See the example:

```
DROME00007    CRAGI20425    n:1
DROME00008    CRAGI20425    n:1
DROME00010    CAPO303580    n:1
NOSA003182    DROME09562    1:m
NOSA003182    DROME09563    1:m
NOSA003195    DROME01960    1:m
```

Then, this orthologous list have been divided in two different files because, as it can be seen in the previous example, in some cases the *D. melanogaster* gene (DROME) is the reference (when it is in the first column, pink lines), while in others it is not (when the *D. melanogaster* gene is in the right column, green lines). So, when the *D. melanogaster* gene is in the first column we only kept those pairs with the tags **n:m** and **n:1**, and if it is in the second column, then we keep **n:m** and **1:m** pairs.


2,635,575 orthologous pairs
D. melanogaster is the reference
D. melanogaster is not the reference
678,777 orthologous pairs
1,954,798 orthologous pairs
271,758 pairs
n:m type: 242,330
n:1 type: 29,428
954,314 pairs
n:m type: 610,316
1:m typ: 343,998
Parsing to obtain paralogous pairs

We parsed the two orthologous pairs lists to obtain the paralogs list by means of a **`bash script`**. The **`bash script`** first take an ortholog gene to the *D. melanogaster* gene (for example, `LUCCU001`) and check how many orthologous genes to *D. melanogaster* has (in the example, `DROME001` and `DROME002`). This indicate that `DROME001` and `DROME002` are paralogous, so they are placed in the same line with an id (in the example, 1) in a **`temporal file`**. The next gene is `LUCCU002`, which has three orthologous *D. melanogaster* genes, `DROME003`, `DROME004` and `DROME005`, so they three are paralogous and are placed in the next line of the **`temporal file`** with the id 2. Finally, the **`bash script`** creates all the possible paralogs combinations from the **`temporal file`**. Normally, they are just pairs of genes (as `DROME001` and `DROME002`), but when there are more than two genes, it creates all the possible paralog pair combinations, as in the case of the paralogs `DROME003`, `DROME004` and `DROME005`, where we obtain three different paralogous pairs.



Orthologous pairs file

```
DROME001 LUCCU001
DROME002 LUCCU001
DROME003 LUCCU002
DROME004 LUCCU002
DROME005 LUCCU002
DROME006 LUCCU003
DROME007 LUCCU003
...
```

Temporal file

```
1:DROME001 DROME002
2:DROME003 DROME004 DROME005
3:DROME006 DROME007
...
```

Final paralog file

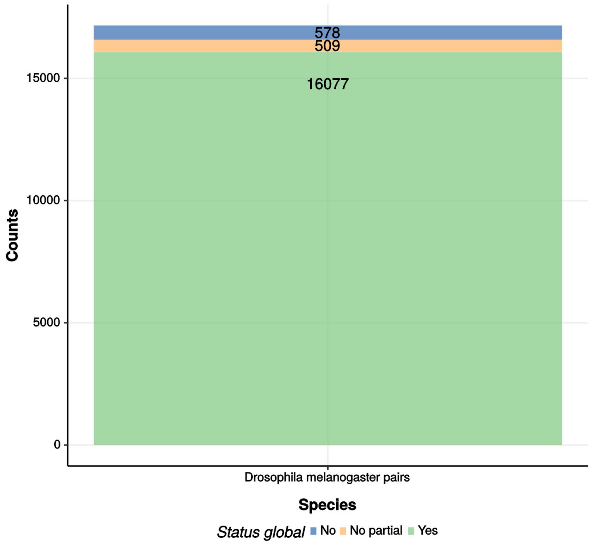
```
1 DROME001 DROME002
2 DROME003 DROME004
3 DROME003 DROME005
4 DROME004 DROME005
5 DROME006 DROME007
...
```

The parsing is done for the two ortholog pairs files that we created. Once we have the two **`final paralogs files`**, we merge them and remove all the duplicated paralogs pairs. Finally, we obtain a total of **17,164 unique paralog pairs**.

2. Domain annotation

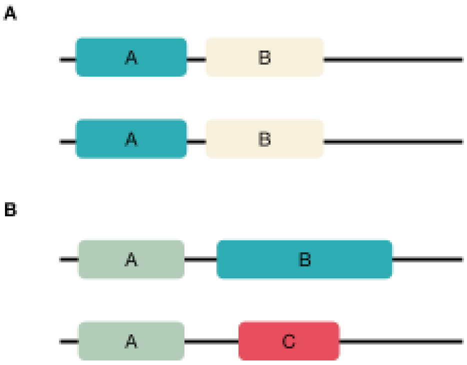
`pfam_scan.pl` is a script that allows users to perform local Pfam searches from FASTA files. We have downloaded the latest version of the script (Pfam 31.0) as well as the latest version of the local libraries and Pfam files to perform the domain annotation. The domain annotation has been performed in the whole *D. melanogaster* proteome available in OMA database, that consist on **14,455 protein sequences**. We have detected 4,358 different domains types in this protein sequences.

From the **17,164 paralog pairs** that we obtained, we can determine that **16,077 pairs** have both genes with at least one domain annotated, in **509 pairs** only one of the genes have a domain annotated ("partial" annotation) and in **578 pairs** none of the genes have a domain annotated. From the pairs with at least one domain annotated in both genes (16,077 pairs), **14,796** have annotated the same number of domains while **1,281** have a different number of domains annotated. We are going to analyze this two cases independently.



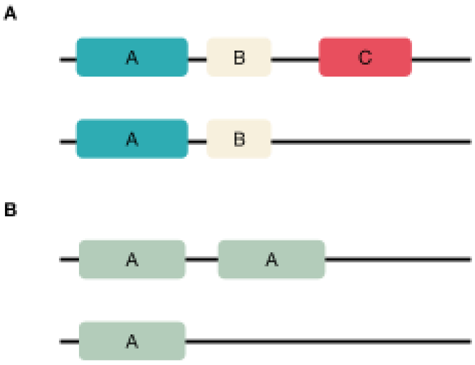
Case 1. Same number of domain annotated

When we have pairs with the same number of domains annotated could be due to: A) they share the same domains or B) some (or all) of the domains are different. For the **14,796 pairs** that have annotated the same number of domains, **14,021 pairs** share the same domains and **775** have different domains. Inside the 14,021 pairs, **14,005 have the same domain order** (a few cases were like A B B - A B A) **We are going to kept the 14,005 pairs that have the same domains annotated in the same order as a "control" for this project**. The hypothesis is that for paralog pairs that share the same domains it is not expected to have different expression patterns, because their functions, as determined by their domains, are going (or expected) to be the same. To clarify, we have not checked the order of the domains, only if they are the same.



Case 2. Different number of domain annotated

If the number of domains is different, it could be due: A) they have different domains annotated or B) one of the genes have a repeated domain. From the **1,281 pairs** with a different number of domains annotated, we find that **980 pairs** are due to a different domain(s) annotated while **301** due to domain repetition. We need to match the 980 pairs with **domainDiff** output to retrieve those pairs that have the domain rearrangements interesting for us: a loss of a domain in the front or in the back of the protein.



3. Domain rearrangement

We run **domainDiff** to determine the domain rearrangements in the Pfam output. The domain rearrangements are classified as:

- Missing one domain in the front (f-1)
- Missing one domain in the back (b-1)
- Ambiguous (f-1|b-1)

domainDiff has detected 35,231 rearrangements in the Pfam output. From them, 24,203 are due to domain loss either in the back or the front of the protein while 11,028 are due to domain repetitions. How many of this 24,203 are truly paralogs according to OMA database? We merge the 980 paralog pair with the **domainDiff** output and we obtain **460 paralogs**.

This 460 paralog pairs are the candidates to be further analyzed with the differential expression analysis to investigate if domain rearrangement play a functional role.

4. Expression analysis (to do)

We are going to use Bgee database and the R package that it provides, **BgeeDB**, to retrieve the expression data for *D. melanogaster*. Another advantage of this database is that is has been recently updated to the version 14-beta (16/05/2017), which include the following new *Drosophila* species:

- *D. ananassae*
- *D. mojavensis*
- *D. pseudoobscura*
- *D. simulans*
- *D. virilis*
- *D. yakuba*

All this species include RNA-seq data (also *D. melanogaster*).

The RNA-seq data for *D. melanogaster* is going to be used to perform the expression analysis in the 460 paralog pairs that we have obtained. On the other hand, in the orthologous analysis that Joaquim is performing, we are going to use the data for the other six species.

The RNA-seq data provided by Bgee is not RPKM, but FPKM. RPKM was made for single-end RNA-seq, where every read correspond to a single fragment that was sequenced. FPKM was made for paired-end RNA-seq. With paired-end RNA-seq, two reads can correspond to a single fragment, or, if one read in the pair did not map, one read can correspond to a single fragment. The only difference between RPKM and FPKM is that FPKM takes into account that two reads can map to one fragment (and so it does not count this fragment twice).

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

We will use the expression breadth as a measure for the tissue or developmental stages specificity. 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 $\tau = 1$, means that is only detectable in one sample (tissue or stage) while $\tau = 0$, that is expressed in all samples with the same expression level.

5. Summary

We have obtained the final candidate list of *D. melanogaster* paralogs with the domain rearrangements that is interesting for us: a total of 460 paralog pairs that have one domain missing either in the back or in the front of the protein. Next, we are going to use Bgee database to study the influence of domain losses over gene expression patterns. This expression patterns are going to be assessed with the expression breadth index as a measure of specificity in either tissues or developmental stages.