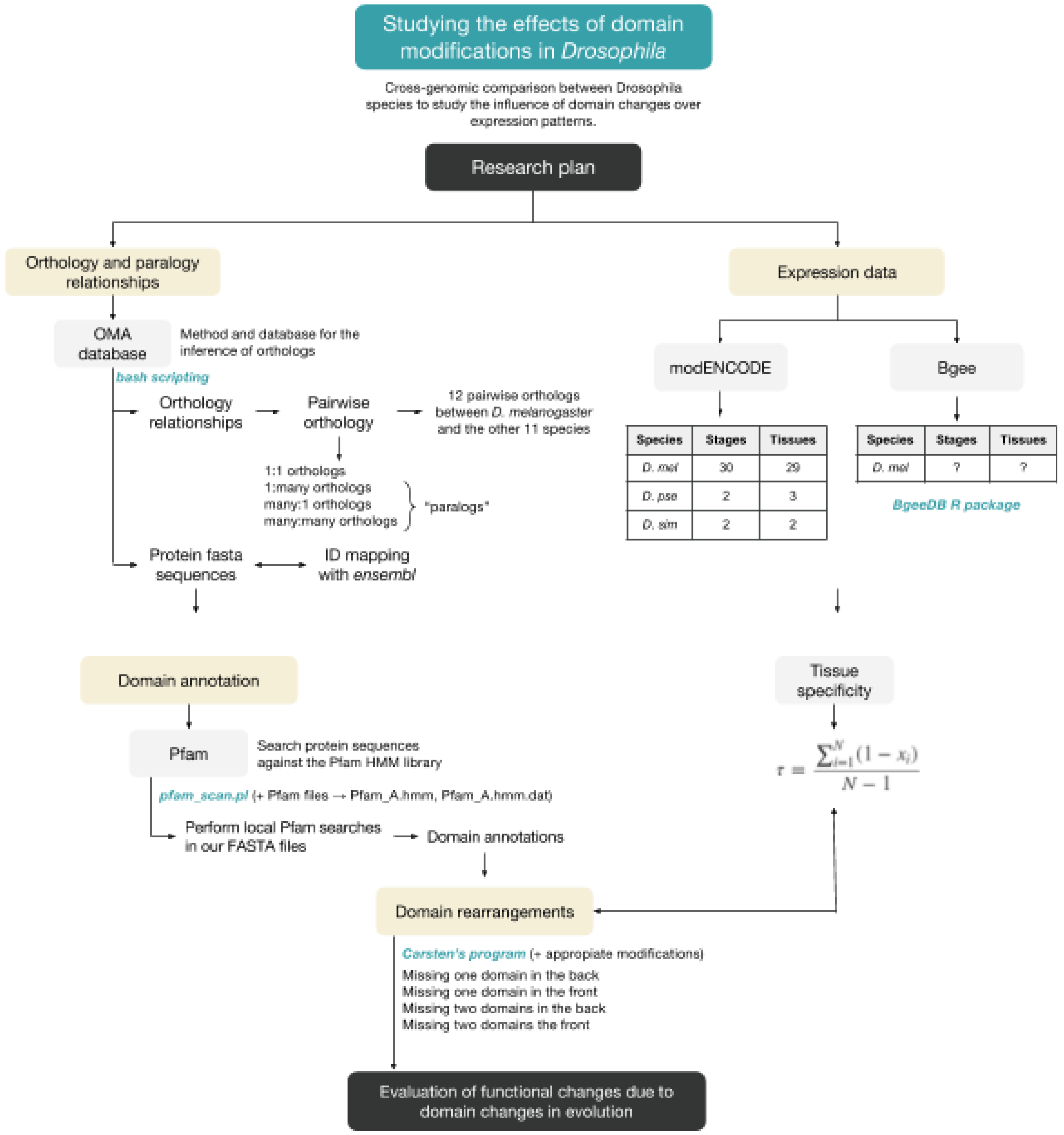
Research plan

The project is divided into the following four main parts:

1. **Orthology and paralogy relationships**. Retrieve the list of orthologs and paralogs candidates between *Drosophila* as well as the protein FASTA sequences of them.
2. **Domain annotation**. The annotation of the domains is going to be done with Pfam.
3. **Domain rearrangement**. Carsten's program in C++ is going to be used to determine the different domain rearrangements.
4. **Expression changes**. To estimate the expression changes we will use Yanai's *et al*. τ index as a measure of tissue or developmental stages bias.


Studying the effects of domain modifications in *Drosophila*
Cross-genomic comparison between Drosophila species to study the influence of domain changes over expression patterns.
Research plan
Orthology and paralogy relationships
OMA database
Method and database for the inference of orthologs
bash scripting
Orthology relationships
Pairwise orthology
12 pairwise orthologs between *D. melanogaster* and the other 11 species
1:1 orthologs
1:many orthologs
many:1 orthologs
many:many orthologs
"paralogs"
Protein fasta sequences
ID mapping with *ensembl*
Expression data
modENCODE
Bgee
BgeeDB R package
Tissue specificity
$\tau = \frac{\sum_{i=1}^{N}(1-x_i)}{N-1}$
Domain annotation
Pfam
Search protein sequences against the Pfam HMM library
pfam_scan.pl (+ Pfam files → Pfam_A.hmm, Pfam_A.hmm.dat)
Perform local Pfam searches in our FASTA files
Domain annotations
Domain rearrangements
Carsten's program (+ appropiate modifications)
Missing one domain in the back
Missing one domain in the front
Missing two domains in the back
Missing two domains the front
Evaluation of functional changes due to domain changes in evolution

Species	Stages	Tissues
D. mel	30	29
D. pse	2	3
D. sim	2	2

Species	Stages	Tissues
D. mel	?	?

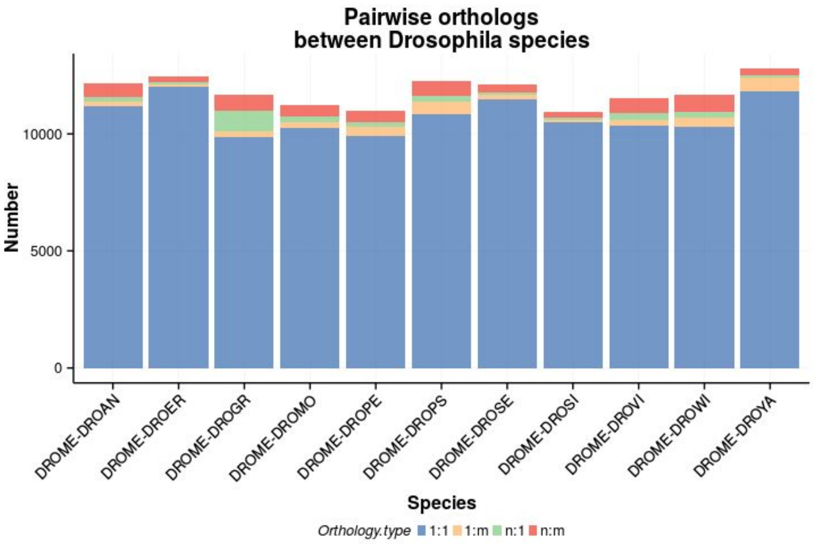
Drosophila species information

We will focus on the *Drosophila* lineage due to the availability of resources, both regarding genomics and expression data. In the OMA database we have, among other Drosophila species, the 12 that conform the *Drosophila* lineage as follows:

OMA IDs	NCBI Taxonomy ID	Species name	OMA data source
DROME	7227	*Drosophila melanogaster*	Ensembl 84; BDGP6 24-FEB-2016
DROAN	7217	*Drosophila ananassae*	Ensembl Metazoa 25 GCA_000005115.1 18-DEC-2014
DROER	7220	*Drosophila erecta*	Ensembl Metazoa 25 GCA_000005135.1 18-DEC-2014
DROGR	7222	*Drosophila grimshawi*	Ensembl Metazoa v5 dgri_r1.3_FB2008_07 15-MAY-2010
DROMO	7230	*Drosophila mojavensis*	Ensembl Metazoa; 21 dmoj_caf1 29-NOV-2013
DROPE	7234	*Drosophila persimilis*	Ensembl Metazoa 25 GCA_000005195.1 18-DEC-2014
DROPS	46245	*Drosophila pseudoobscura*	Ensembl Metazoa 21 HGSC2 29-NOV-2013
DROSE	7238	*Drosophila sechellia*	Ensembl Metazoa 23; dsec_caf1 24-JUL-2014
DROSI	7240	*Drosophila simulans*	Ensembl Metazoa 25 GCA_000259055.1 18-DEC-2014
DROVI	7244	*Drosophila virilis*	Ensembl Metazoa v5 dvir_r1.2_FB2008_07 15-MAY-2010
DROWI	7260	*Drosophila willistoni*	Ensembl Metazoa 3 dwil_r1.3_FB2008_07 20-AUG-2009
DROYA	7245	*Drosophila yakuba*	Ensembl Metazoa 21 dyak_r1.3_FB2008_07 29-NOV-2013

Pairwise orthologs between Drosophila species

The pairwise orthologs between Drosophila species have been retrieved from the OMA database. The following graph summarize the pairwise orthologs between *Drosophila melanogaster* and the other 11 species inside their phylogeny. In all the cases, the majoritary type of **orthology** is the **1:1 type**. It can be considered that the other three types: **1:m**, **n:1** and **n:m** are **paralogous** genes as they imply a duplication event at some point of the divergence process.



The following table summarizes the number of orthologous between the species, the number of paralogous and the number of FASTA sequences for the paralogs.

Pair group	Number of orthologous (1:1, 1:m, n:1, n:m)	Number of paralogous (1:m, n:1, n:m)	Paralogous FASTA sequences
DROME DROAN	12154	972	449 455
DROME DROER	12478	464	234 214
DROME DROGR	11666	1801	844 1197
DROME DROMO	11231	994	540 537
DROME DROPE	10993	1097	589 653
DROME DROPS	12280	1457	713 846
DROME DROSE	12131	647	251 348
DROME DROSI	10932	458	241 250
DROME DROVI	11536	1189	570 590
DROME DROWI	11666	1377	640 735
DROME DROYA	12786	985	495 762

Pfam: searching domains

Poteomes are going to be scanned using the **pfamscan** utility, specifically, the script they provide that allows users to perform local Pfam searches: `pfam_scan.pl`.
The information that we need is the hmm acc, that is the ID of the domain, and the start and end of the domain.

Determine domain rearrangements

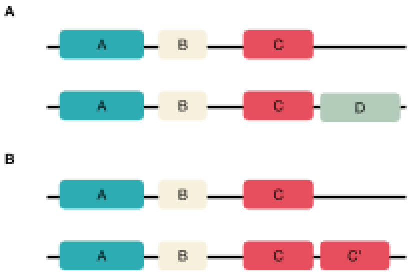
The protein domain rearrangements can be classified as:

- Missing one domain in the front
- Missing two domains in the front
- Missing one domain in the back
- Missing two domains in the back
- Missing one domain the front and one in the back

The consideration of **homologs without domain modifications** can be used as a “control”, and compare them against paralogs with a domain rearrangement.

We will use **Carsten’s program (C++ script)** to determine the domain rearrangements commented above. We will apply all the changes necessaries to fit the program to our input data and the output that we need.

Check: we need to ensure that the domains differences between two paralogs are not due to the repetition of domains (as happens in B). In this case, it is expected that the function of the two paralogs would still be the same or similar. We are interested in cases as A, where a new domain is gained.



Measure of tissue/stages specificity

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 RPKM 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.

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