M1 or M2 internship: Analysis of the genomic structure of *Drosophila melanogaster* : is the evolution of duplicated genes related with their environment in transposable elements? (Version française ci-dessous)

Co-supervisor: Emmanuelle Lerat
Coordonnées: emmanuelle.lerat(AT)univ-lyon1.fr
Laboratory « Biométrie et Biologie Évolutive » UMR5558 University Lyon 1

Co-supervisor: Carène Rizzon
Coordonnées: carene.rizzon(AT)univ-evry.fr
Laboratory "Mathématiques et Modélisation d'Évry" UMR8071 University of Evry

Training place: Lab. Biométrie et Biologie Evolutive (LBBE), 43 boulevard du 11 novembre 1918, Université Lyon 1 Campus La Doua, 69622 Villeurbanne cedex **et/ou** Laboratoire de Mathématiques et Modélisation d'Évry (LaMME) UMR8071 Université d'Evry, bat. IBGBI, 23 Bd de France, 91037 Evry CEDEX.

Within genomes, duplicated genes (paralogs) are formed by different mechanisms such as the complete duplication of the whole genome, the action of transposable elements (TEs), segmental duplications, and tandem duplications (1). These genes, after duplication, can be subjected to various evolutionary processes allowing their maintenance or their loss (acquisition of a new function (neo-functionalization), sharing of the ancestral function (sub-functionalization), pseudogenization, functional redundancy by dosage effect) (1). Duplicated genes constitute families of genes and are of great importance in the formation of new genes and in creating genetic novelty in organisms. Many new gene functions have evolved through this mechanism. However, the processes allowing the maintenance of duplicated genes within genomes remain poorly understood. In particular, little is known about the influence of TEs at this level.

TEs are repeated sequences that have the ability to move within the genome. They are now recognized as having a significant impact on the evolution of genomes and the adaptation of species (2). In the model species *Drosophila melanogaster*, it has been shown that duplicated genes constitute around 40% of all genes (1), the majority of which are thought to be the result of tandem duplications (3). The most recent duplicated genes seem more often subject to the neo-functionalization mechanism (4) and their functions are mainly linked to responses to environmental stresses (5). Within this genome, we find around 15% of TEs, the distribution of which is not random (6). **We can therefore wonder about the importance of TEs in the maintaining of the different families of genes in this species.**

Required skills: This internship involves performing analyses of NGS and genomic data using various existing tools and using statistical tests. It is therefore necessary to have very good notions in statistics (R software), skills in Python and shell programming, but also a particular interest in molecular evolution.

The goal of the internship: The internship aims to better understand the evolution of duplicated genes in *Drosophila melanogaster* by taking into account the environment in transposable elements (TEs) present around and in the genes. *D. melanogaster* is a model organism for which numerous data of a varied nature and used in comparative genomics are available. This internship follows on from several studies which have made it possible to

determine the families of duplicated genes and the ET environment of all the genes in Drosophila.

1) Determination of unambiguous duplications
For each family of duplicated genes (sizes 3 to 10 with a dataset already created in the laboratory), we will construct a phylogenetic tree in order to select the most recent pairs of duplicated genes, therefore representing unambiguous duplications.

2) Taking into account size 2 families
By including gene families of size 2, we will test the hypothesis that pairs of genes from the same duplication event tend to have a more similar TE environment than random gene pairs. The age of the duplication (estimated via the calculation of the synonymous substitution rate (Ks)) and the position of the duplicated genes relative to each other will be taken into account in this analysis.

3) Identification of pseudogenes resulting from duplications
From the pseudogenes annotated in the Drosophila genome and by comparison of sequences (using BLAST), we will identify the pseudogenes resulting from ancestral duplication events. This data will allow a detailed study of the question of what types of genes (in terms of function, location and environment in TEs) are maintained in the genome after duplication.

4) Analysis of the expression of duplicated genes
Using the pairs of duplicated genes determined in points 1 and 2, we will estimate the variation of expression according to the environment in TEs and to the Ks under several conditions (head, ovary, larva, embryo) in order to estimate the influence of each parameter on this variation.

5) Functional conservation of duplicated genes
Using “gene ontology” data and data from protein-protein interaction networks, we will test whether the functional conservation between pairs of duplicated genes is associated with the presence of TEs near these genes, taking into account the age of the duplications.

References

1. Zhang. Trends Genet. 2003
2. Bourque et al. Genome Biol. 2018
3. Zhou et al. Genome Res. 2008
4. Assis and Bachtrog. PNAS 2013
5. Zhong et al. BMC genomics 2013
6. Adams et al. Science 2000

Sujet stage M1 ou M2 2021-2022 : Analyse de la structure du génome de *Drosophila melanogaster* : l'évolution des gènes dupliqués est-elle influencée par leur environnement en éléments transposables ?

Co-Responsable : Emmanuelle Lerat
Coordonnées : emmanuelle.lerat(AT)univ-lyon1.fr
Laboratoire « Biométrie et Biologie Évolutive » UMR5558 Université Lyon 1

Co-Responsable : Carène Rizzon
Coordonnées : carene.rizzon(AT)univ-evry.fr
Laboratoire de Mathématiques et Modélisation d'Évry UMR8071 Université d'Evry

Lieu du stage: Lab. Biométrie et Biologie Evolutive (LBBE), 43 boulevard du 11 novembre 1918, Université Lyon 1 Campus La Doua, 69622 Villeurbanne cedex **et/ou** Laboratoire de Mathématiques et Modélisation d'Évry (LaMME) UMR8071 Université d'Evry, bat. IBGBI, 23 Bd de France, 91037 Evry CEDEX.

Au sein des génomes, les gènes dupliqués (paralogues) sont formés par différents mécanismes tels que la duplication complète du génome, l'action d'éléments transposables (ET), les duplications segmentales, et les duplications en tandem (1). Ces gènes, après duplication, peuvent être soumis à différents processus évolutifs permettant leur maintien ou leur perte (acquisition d'une nouvelle fonction (néo-fonctionalisation), partage de la fonction ancestrale (sous-fonctionalisation), pseudogénisation, redondance fonctionnelle par effet de dosage) (1). Les gènes dupliqués constituent des familles de gènes et ont une grande importance dans la formation de nouveaux gènes et pour créer de la nouveauté génétique dans les organismes. Beaucoup de nouvelles fonctions de gènes ont évolué grâce à ce mécanisme. Cependant, l'ensemble des processus permettant le maintien des gènes dupliqués au sein des génomes reste mal connu. En particulier, on connaît mal l'influence des ET à ce niveau.

Les ET sont des séquences répétées qui ont la capacité de se déplacer dans les génomes. Ils sont maintenant reconnus comme ayant un impact important dans l'évolution des génomes et l'adaptation des espèces (2). Chez l'espèce modèle *Drosophila melanogaster*, il a été montré que les gènes dupliqués constituent de l'ordre de 40 % de l'ensemble des gènes (1) dont la majorité serait issue de duplications en tandem (3) . Les gènes dupliqués les plus récents semblent plus souvent soumis au mécanisme de néo-fonctionnalisation (4) et leurs fonctions seraient principalement reliées aux réponses à des stress environnementaux (5). Au sein de ce génome, on trouve de l'ordre de 15% d'ET dont la répartition n'est pas aléatoire (6). **On peut donc se poser la question de savoir quelle est l'importance des ET dans le maintien des différentes familles de gènes dans cette espèce.**

Compétences requises

Ce stage implique d'effectuer des analyses de données NGS et génomiques en utilisant différents outils existants et d'utiliser des tests statistiques. Il faut donc avoir de très bonnes notions en statistiques (logiciel R), des compétences en programmation Python et shell, mais aussi un intérêt particulier dans l'évolution moléculaire.

Le but du stage

Le stage vise à mieux comprendre l'évolution des gènes dupliqués chez *Drosophila melanogaster* en tenant compte de l'environnement en éléments transposables (ET) présents autour et dans les gènes. *D. melanogaster* est un organisme modèle pour lequel de nombreuses données de nature variée et utilisées en génomique comparative sont disponibles. Ce stage fait suite à plusieurs travaux qui ont permis de déterminer les familles de gènes dupliqués et l'environnement en ET de l'ensemble des gènes chez la drosophile.

1) Détermination des duplications non ambiguës

Pour chaque famille de gènes dupliqués (taille 3 à 10 avec un jeu de données déjà constitué au laboratoire), il s'agira de construire un arbre phylogénétique afin de sélectionner les paires de gènes dupliqués les plus récentes, représentant donc des duplications non ambiguës.

2) Prise en compte des familles de taille 2

En incluant les familles de gènes de taille 2, on testera l'hypothèse que les paires de gènes issus d'un même événement de duplication ont tendance à avoir un environnement en ET plus similaire que des paires de gènes aléatoires. L'âge des duplications (estimé via le calcul du taux de substitution synonyme (Ks)) et la position des gènes dupliqués l'un par rapport à l'autre seront pris en compte dans cette analyse.

3) Identification des pseudogènes issus de duplications

A partir des pseudogènes annotés dans le génome de la drosophile et par comparaison de séquences (en utilisant BLAST), nous identifierons les pseudogènes issus d'événements ancestraux de duplications. Ces données permettront d'étudier de manière fine la question de quels types de gènes (en termes de fonction, de localisation et d'environnement en ET) sont maintenus dans le génome après duplication.

4) Analyse de l'expression des gènes dupliqués

En utilisant les paires de gènes dupliqués déterminés aux points 1 et 2, nous estimerons la variation d'expression en fonction de l'environnement en ET et du Ks dans plusieurs conditions (tête, ovaire, larve, embryon) afin d'estimer l'influence de chaque paramètres sur cette variation.

5) Conservation fonctionnelle des gènes dupliqués

En utilisant des données de "gene ontology" et des données de réseaux d'interaction protéine-protéine, nous testerons si la conservation fonctionnelle entre paires de gènes dupliqués est associée à la présence d'ET près de ces gènes, en tenant compte de l'âge de la duplication.

Références

1. Zhang. Trends Genet. 2003
2. Bourque et al. Genome Biol. 2018
3. Zhou et al. Genome Res. 2008
4. Assis and Bachtrog. PNAS 2013
5. Zhong et al. BMC genomics 2013
6. Adams et al. Science 2000