Quantification non-biaisée du fardeau génétique dans des super-gènes de plantes

(english version below)

Université Claude Bernard Lyon 1
Laboratoire de Biométrie et Biologie Evolutive UMR CNRS 5558 Villeurbanne

Encadrante : Hélène Badouin (*helene.badouin@univ-lyon1.fr*), ==
https://lbbe.univ-lyon1.fr/-Equipe-Sexe-et-Evolution-.html?lang=fr,
https://scholar.google.fr/citations?user=QkgCKZsAAAAJ&hl=fr

Contexte :

Certains génomes eucaryotes possèdent des régions multigéniques qui échappent à la recombinaison méiotique : c'est ce qu'on appelle les "super-gènes". Les chromosomes sexuels qui contrôlent le sexe génétique en sont l'exemple le plus connus, mais des super-gènes contrôlent aussi des phénotypes tels que le déterminisme social chez la fourmi de feu, le mimétisme chez les papillons Heliconus, ou l'incapacité de nombreuses plantes à fleur de s'autoféconder (auto-incompatibilité). Ces super-gènes possèdent plusieurs versions (haplotypes) parfois très divergentes entre elles, **qui ne recombinent pas entre elles et sont maintenues au cours du temps par de la sélection équilibrante**. A cause de l'absence de recombinaison entre haplotypes, ceux-ci accumulent des mutations qui peuvent être délétères (**notion de "fardeau génétique"**). De plus, les haplotypes minoritaires (par exemple, le Y dans les systèmes XY) ont un fardeau génétique plus lourd, ce qui peut conduire à une perte d'expression ou de séquence plus rapide. Des **mécanismes de compensation** ont évolué dans certaines espèces, par exemple la compensation de la diminution d'expression d'un gène Y par une plus forte expression de la copie X, appelée compensation de dosage. Afin de caractériser et de quantifier les phénomènes de fardeau génétique et de compensation, il est nécessaire de **reconstituer et comparer précisément les séquences des différents haplotypes**. Or, cette reconstitution est souvent partielle et biaisée avec les méthodes actuelles.

But du stage :

Notre équipe a développé des méthodes permettant d'identifier des polymorphismes XY en étudiant leur transmission ou leur répartition dans un ensemble d'individu (Muyle et al GBE 2016, Käfer et al. in prep). J'ai mis au point une méthode de **reconstruction bio-informatique ("micro-assemblage") des différents haplotypes de super-gènes** à partir de ces polymorphismes et des données brutes qui ont permis de les détecter. L'objet de ce stage est d'appliquer cette méthode à des jeux de données sur plusieurs super-gènes pour **améliorer l'estimation du fardeau génétique et estimer si/à quel point les méthodes précédentes introduisaient des biais**.

Pour cela, l'étudiant ou étudiante disposera de plusieurs jeux de données de transcriptomique. Pour ces jeux de données, la séquence complète de certains haplotypes est déjà obtenue et validée expérimentalement, ce qui permettra de valider les résultats du micro-assemblage.

L'analyse consistera en les étapes suivantes :

1 Reconstruire les séquences d'haplotypes même très divergents.
2 Aligner les séquences des différents haplotypes.
3 Calculer des statistiques de génétique des populations pour quantifier le fardeau génétique global et/ou par haplotype, par exemple en calculant le ratio de la diversité non synonyme sur la diversité synonyme (π_n/π_s).
4 Mesurer l'expression de chaque haplotype pour détecter une éventuelle compensation de dosage.

Compétences : des compétences en bio-informatique (utilisation d'un environnement Unix, d'un cluster informatique partagé, langage script Python ou Perl, R) sont très fortement recommandées. Le stage permettra d'acquérir des compétences en analyse de données NGS et génomique des populations.

Sélection de publications :

Badouin, H. et al. (2017) *The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution*. Nature 546, 148–152.

Branco, S.*, **Badouin, H.***, Rodríguez de la Vega*, R.C., Gouzy, J., Aguileta, G., Siguenza, S., Brandebourg, J.-T., Coelho, M.A., Hood, Giraud, T (2017). *Evolutionary strata on young mating-type chromosomes despite lack of sexual antagonism*. PNAS 114, 7067-7072. doi: 10.1073/pnas.1701658114

Muyle, A., **J. Käfer**, N. Zemp, S. Mousset, F. Picard et G. A. Marais (2016). *SEX-DETector: A Probabilistic Approach to Study Sex Chromosomes in Non-Model Organisms*. Genome Biol. Evol. 8.8, p. 2530-2543.

Muyle, A., Zemp, N., Fruchard, C., Cegan, R., Vrana, J., Deschamps, C., ... & Marais, G. A. (2018). *Genomic imprinting mediates dosage compensation in a young plant XY system*. Nature plants, 4(9), 677.

Raymond, O.*, Gouzy, J.*, Just, J.*, **Badouin, H.***, Verdenaud, M.*, et al. (2018). *The Rosa genome provides new insights into the domestication of modern roses.* Nature Genetics 50(6), 772-777.

Unbiased quantification of genetic burden in plant supergenes

Some eukaryotic genomes have multigene regions that escape meiotic recombination: these are called "super-genes". Sex chromosomes that control genetic sex are the best-known examples, but super-genes also control phenotypes such as social determinism in fire ant, mimicry in Heliconus moths, or the inability of many flowering plants to self-fertilize (self-incompatibility). These super-genes have several versions (haplotypes) sometimes very divergent, which do not recombine with each other and are maintained over time by balancing selection. Because of the absence of recombination between haplotypes, they accumulate mutations that can be deleterious (notion of "genetic burden"). In addition, minority haplotypes (for example, the Y in XY systems) have a heavier genetic burden, which can lead to a faster loss of expression or sequence. Compensation mechanisms have evolved in some species, for example, compensating the decreased Y expression of a gene by a higher expression of the X copy, a phenomenon called "dosage compensation". In order to characterize and quantify the genetic burden and compensation phenomena, it is necessary to reconstitute and compare precisely the sequences of the different haplotypes. However, this reconstitution is often partial and biased with current methods.

Our team has developed methods for identifying XY polymorphisms by studying their transmission or distribution in a set of individuals (Muyle et al GBE 2016, Käfer et al., In prep). I developed a method of bioinformatic reconstruction ("micro-assembly") of different haplotypes of super-genes from these polymorphisms and the raw data that allowed to detect them. **The goal of this training is to apply this method to datasets on several super-genes to improve the estimation of the genetic burden and to estimate if / to what extent the previous methods introduced biases**.

For this purpose, the student will have several sets of transcriptomic data. For these datasets, the complete sequence of some haplotypes is already obtained and validated experimentally, which will allow validating the results of the micro-assembly.

The analysis will consist of the following steps:

1- Rebuild haplotype sequences even if they are highly divergent.

2- Align the sequences of the different haplotypes.

3- Calculate population genetics statistics to quantify the overall and / or haplotype genetic burden, for example by calculating the ratio of non-synonym diversity over synonymous diversity ($\pi n / \pi s$).

4- Measure the expression of each haplotype to detect possible dosage compensation.

Bioinformatics skills (using a Unix environment, a shared IT cluster, Python or Perl scripting language, R) are highly recommended. The internship will acquire skills in NGS data analysis and genomic populations.