Large-scale assessment of the impact of protein sequence variations on ageing using *Drosophila melanogaster*

Background. The growing body of available genomics and transcriptomics data has opened exciting research avenues toward characterizing the relationship between sequence variations and phenotypes. However, the number of variants observed in a given population generally largely exceeds the number we are able to investigate experimentally. To address this issue, several computational methods massively exploiting the information encoded in natural protein sequences have been recently proposed [1,2,3]. They predict the extent to which a mutation, a combination of mutations or an indel is likely to impair a protein's function. They have greatly improved the prediction of mutational phenotypic outcomes, when assessed against deep mutational scans [1,2,3] and clinical annotations [4]. Alternatively, Genome Wide Association Studies (GWAS) are a powerful tool to investigate the impact of genetic variations, especially when experimental measures of a particular phenotype across a population are available [5].

Objectives. In this project, we propose to combine high-throughput computational scans with *in vivo* experiments to identify and validate a set of genes and associated sequence variations involved in ageing. We will use Drosophila as our model system, and our analysis will be performed at the scale of its protein-coding genome. We will use the Smurf phenotype as our indicator of life expectancy in flies [6].The specific objectives of the project are the following:

(1) leverage the massive amount of available sequence data to provide an extensive and comprehensive resource quantifying the effects of all possible substitutions at all positions of the Drosophila proteome.
(2) Identify a set of non-synonymous single nucleotide polymorphism (SNPs) highly relevant for ageing and validate them by *in vivo* experiments on flies.
(3) Build a predictive and interpretable model to characterize the way alternative splicing shapes Drosophila's proteome and interactome during ageing.

Data. We will predict the full single-mutational landscapes of all the proteins annotated in FlyBase [7]. For each protein, we will retrieve a set of related sequences by searching three public databases: UniRef90 [8], BFD [9-10] and MGnify clusters [11]. They include annotated protein sequences and also sequences coming from metagenomics experiments. We estimate the number of retrieved sequences to be between a few tens and several millions depending on the protein. We will retrieve population-wide genomic sequence variations from the Drosophila Genetic Resource Panel (DGRP) [12]. This community resource provided annotations for the genetic polymorphisms observed in a panel of 200 inbred lines. We will exploit transcriptomic data coming from two sources, public databases such as Ensembl [13] (gene annotations) and Bgee [14] (RNA-Seq splice junctions), and our own experiments. Finally, we will exploit longevity data we collected on 117 fly lines from DGRP.

Methodology. To predict full single-mutational landscapes, we will rely on GEMME[1] [3], an efficient method developed by EL, which explicitly models inter-dependencies between protein residues to predict mutational outcomes. Compared to other state-of-the-art methods, GEMME is very fast, does not require training, contains a very small number of parameters, and performs well even in extreme conditions, where the input sequence variability is low. The predictions computed by GEMME and other

[1] http://www.lcqb.upmc.fr/GEMME

related methods are not specific to a particular phenotype, but they are good indicators of whether a mutation or combination of mutations is likely to impair a protein's function. The candidate will systematically compare the mutational outcomes predicted by GEMME with results coming from a GWAS analysis based on experimental measurements of the Smurf phenotype [12,13] MR's team collected on 117 DGRP lines. By using a novel theoretical framework, the 2-phase model of ageing [12], MR's team was able to detect a strong association signal for ageing (Fig. 1, left panel). The SNPs of interest identified with this approach will allow narrowing down the number of genes to target in order to test their role in ageing experimentally. The candidate will functionally validate the candidate genes using the drosophila model and conditional down/up-regulation thanks to the gene switch system. In addition, we plan to engineer the genome of one of the short-lived DGRP line by replacing some of the strongest candidates our approach will identify using a CRISPR-Cas9 approach. To investigate the implication of alternative splicing in ageing, we will rely on ThorAxe[2] [15] and PhyloSofs[3] [16], a couple of efficient methods recently developed in EL's team to assess the evolutionary conservation of splice variants and predict their 3D structures. The candidate will systematically quantify the evolutionary conservation of splice variants whose expression varies with age in Drosophila. Her/His working hypothesis will be that variants conserved in evolution are likely to be functionally relevant. This will help her/him to filter out variants likely associated with "noise" and to identify a set of variants whose function is modulated during lifespan and/or whose function modulation is associated with the end-to-life process. S/he will also assess the impact of the variations on the 3D structures of the corresponding proteins and on their interactions. To this end, s/he will develop a method to extract co-variation patterns of alternative splicing between two or more protein partners.

Preliminary work. During Spring 2020 (COVID-related lockdown), we managed to make significant progress thanks to a master student we (EL and MR) co-supervised together. This led to the development of a computational tool for computing and visualizing a GEMME-predicted mutational outcome profile for any drosophila protein, along with GWAS statistical p-values (for any phenotype) and allele frequencies using the DGRP lines (Fig. 1, right panel).

Expected outcomes

1. An optimized version of GEMME allowing full genome scans.
2. An online database connected to FlyBase in order to implement this new theoretical result into that broadly used drosophila online resource.
3. A list of SNPs with a highly plausible role in ageing and functional validation of the associated genes.
4. An end-of-life alternative splicing profile of drosophila females with the associated polymorphisms of splicing sites.
5. An edited fly genome to test the functional role of given SNPs in ageing.

Interdisciplinarity. This project is at the cross-talk between genomics/transcriptomics, evolution, structural bioinformatics and *in vivo* experimental biology. The candidate will share her/his time between two groups located at the LCQB (IBPS, SU-CNRS) and the CRI Paris. Both environments are highly interdisciplinary and will provide her/him the resources needed to address the hypothesis proposed in this project (high-throughput computing facilities, production of CRISPR-based massive genome editing approach facilities). Under the supervision of Elodie Laine at the LCQB, the candidate will benefit from a strong expertise in computational biology. Under the supervision of Michael Rera at CRI, the candidate will benefit from knowledge and know-how about the biology of ageing in Drosophila.

[2] https://github.com/PhyloSofS-Team/thoraxe
[3] https://github.com/PhyloSofS-Team/PhyloSofS

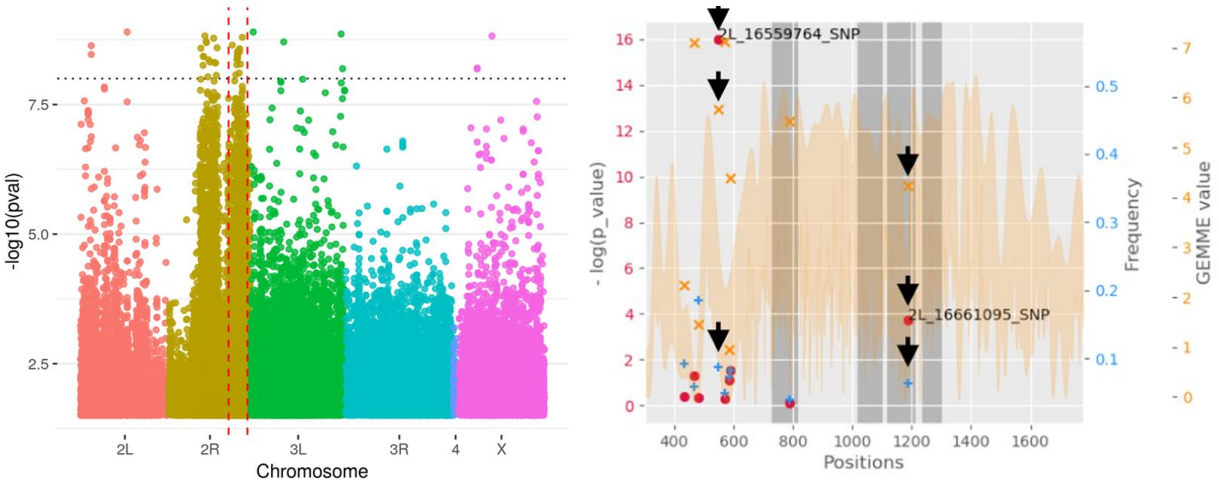

Figure 1: Left. Optimization of the parameters used for ageing GWAS. We used 117 lines from the DGRP set with lifespans ranging from 15 to 67 days. By changing the parameter from the mean lifespan from the rate of Smurfs appearance in the populations, we managed to identify a genomic inversion as being associated with an average 50% lifespan increase, and globally improved the statistical association between some SNPs and the phenotype. **Right. Proteome-wide profiling in *D. melanogaster.*** We developed a tool that evaluates the GEMME score of each DGRP SNP for any given drosophila protein. This tool allows the profiling of GEMME scores along the amino acid sequence and locates SNPs that were identified in any GWAS approach, facilitating the identification of SNPs with a plausible function in our phenotype, here ageing.

Expected profile. The PhD candidate should have strong programming skills and experience with high-throughput computing. Previous experience dealing with evolutionary aspects and/or protein sequence analysis is a plus. S/he should show some interest in molecular biology. The project taking place in two distinct laboratories, the candidate should have some capacity to adapt to different environments and should be able to communicate easily with people coming from different backgrounds. In addition, our approaches being strongly interdisciplinary, we are searching for someone highly interested by multiple research domains including mathematics and biology.

References

1. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. (2016) *Mol Biol Evol*. 33(1):268–280.
2. Riesselman AJ, Ingraham JB, Marks DS. (2018) *Nat Methods*. 15(10):816–822.
3. Laine, E., Karami, Y., & Carbone, A. (2019). *Molecular biology and evolution*, *36*(11), 2604-2619.
4. Frazer, J., Notin, P., Dias, M., Gomez, A., Brock, K., Gal, Y., & Marks, D. (2020). *bioRxiv*.
5. Tricoire, H. & Rera, M. A (2015) *PLoS One* 10, e0141920 .
6. Rera, M., Clark, R. I. & Walker, D. W. (2012) *Proc Natl Acad Sci USA* 109, 21528–33.
7. Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati PV, Goodman JL, Gramates LS, Millburn G, Strelets VB, Tabone CJ, Thurmond J and the FlyBase Consortium (2021) *Nucleic Acids Res*. 49(D1) D899–D907
8. UniProt Consortium. (2019). *Nucleic Acids Research*, 47(D1), D506-D515.
9. Steinegger, M., Mirdita, M., & Söding, J. (2019). *Nature Methods*, 16(7), 603-606.
10. Steinegger, M., & Söding, J. (2018). *Nature Communications*, 9(1), 1-8.
11. Mitchell, A. L. et al. (2020). *Nucleic Acids Research*, 48(D1), D570-D578.
12. Mackay, T. F. *et al.* (2012) *Nature* 482, 173–8.
13. Yates, A. *et al.* (2016) *Nucleic Acids Res.*, 44, D710–716.
14. Komljenovic, A.; Roux, J.; Wollbrett, J.. *peer review: 2 approved, 1 approved with reservations* 2018,
15. Zea, D. J., Laskina, S., Baudin, A., Richard, H., Laine, E.. (2020). *bioRxiv*.
16. Ait-hamlat, A.; Zea, D. J.; Labeeuw, A.; Polit, L.; Richard, H.; Laine, E. (2020). *J Mol Biol* 432(7):2121-2140