Développement et évaluation de méthodes statistiques d'inférence de la sélection à partir de données génomiques temporelles
(Developing and evaluating statistical inference methods for detecting selection from temporal series of genomic data)

PhD by Cyriel Paris
Defended on March 10, 2020, at INP Toulouse, France, for the SEVAB doctoral school

Because the genetic diversity of a population is shaped by its evolutionary history, it can be used to infer some aspects of this history, for instance it can lead to detect the genomic regions under selection. Inference about past selection is generally based on genomic data from contemporary individuals and looks for specific patterns of genetic diversity in these data. With the recent advances of sequencing technologies, collecting genomic samples at several dates in the same population has also become possible. Such temporal data provide direct access to the past evolution of genetic diversity, which can be exploited for selection inference. However, taking advantage of this information requires to develop new methods dedicated to the analysis of genomic time series. This question is commonly tackled by the use of a hidden Markov model (HMM), which allows to exploit allele frequency evolution while accounting for the sampling noise associated to observed allele frequencies. However, one key question in this context is how to model the stochastic evolution of allele frequencies. While the Wright-Fischer model with selection is a natural choice, computing the likelihood of an observed trajectory under this model is computationally prohibitive.
Therefore, several approximations of this process have been proposed, based either on the resolution of a differential equation (the diffusion equation) or on the use of usual parametric distributions whose moments match those of the Wright-Fischer. During this PhD, I first studied a method based on an elegant analytical resolution of the diffusion equation, but this method was found very difficult to use in practice due to high computational costs and numerical instability issues. Next, I compared several parametric distributions, including the Beta with Spikes distribution.
Intially proposed for neutral models, this distribution was considered for the study of selection by Tataru et al. (2017). The originality of this model is to account for fixation probabilities, that is for the probability that one allele gets fixed or lost during evolution. I included this distribution into a HMM framework and showed that it was a good approximation of the Wright-Fisher process, which lead to accurate selection inference. Finally, I demonstrated the feasability and the interest of this approach by analyzing three real time series with high density genomic data, which lead to the detection of several genomic regions under selection. I also analyzed these data using standard methods based on a single present time sample and found quite different signals. This outlines the complementarity of time series and present time approches, which can be combined for a better global understanding of evolutionary history.