Selnetime: A new method inferring demography and selection from genomic time series data

Mathieu Uhl, Simon Boitard, Miguel De-Navascues-Melero, Bertrand Servin

SELNETIME : A NEW METHOD INFERRING DEMOGRAPHY AND SELECTION FROM GENOMIC TIME SERIES DATA

Mathieu Uhl [1], Simon Boitard [2], Miguel De-Navascues-Melero [2], Bertrand Servin [3]
[1] CEFE, CNRS , [2] CBGP, INRAE, [3] GenPhySE, INRAE

mathieu.uhl@cefe.cnrs.fr, simon.boitard@inrae.fr, miguel.navascues@inrae.fr, bertrand.servin@inrae.fr

Introduction

Traditionally, contemporary individuals are sampled to determine demographic patterns or regions under selection. However, genetic data can be affected by various confounding factors, including both selection and demography. To mitigate these effects, time series data can be used. These data, which can be obtained from monitoring natural populations, experimental populations or ancient DNA, have several applications. They are valuable for population management, for understanding the genetic basis of traits of interest, and for characterising evolutionary history.

Material and methods

Evolution of alleles frequency

The evolution of allele frequencies over time provides critical insights into both demography and selection. The trajectory of these frequencies reveals how populations evolve and adapt.

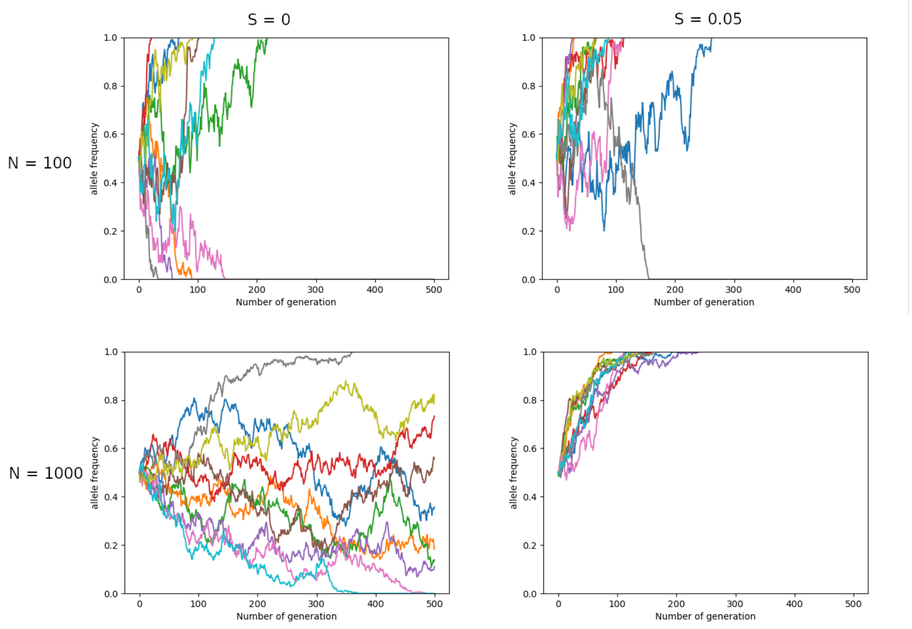

Figure 1: Information on demography and selection

Information about selection can be inferred from the direction of the frequency changes: alleles under positive selection tend to increase in frequency, while those under negative selection decrease. On the other hand, demographic information is in the variance of the allele frequency trajectory. In larger populations, the variance is typically lower, whereas smaller populations exhibit higher variance because genetic drift has a more pronounced effect. Thus, by analyzing both the direction and variance of allele frequency trajectories, we can disentangle the influences of selection and demographic factors on genetic evolution.

Hidden Markov model (HMM) (1)

This approach allows to precisely infer the hidden states of evolutionary processes by analyzing observable genetic data. However, the application of HMMs requires several stong hypothesis about the evolutionary dynamics of the populations under study. These assumptions typically include panmixia and isolation.

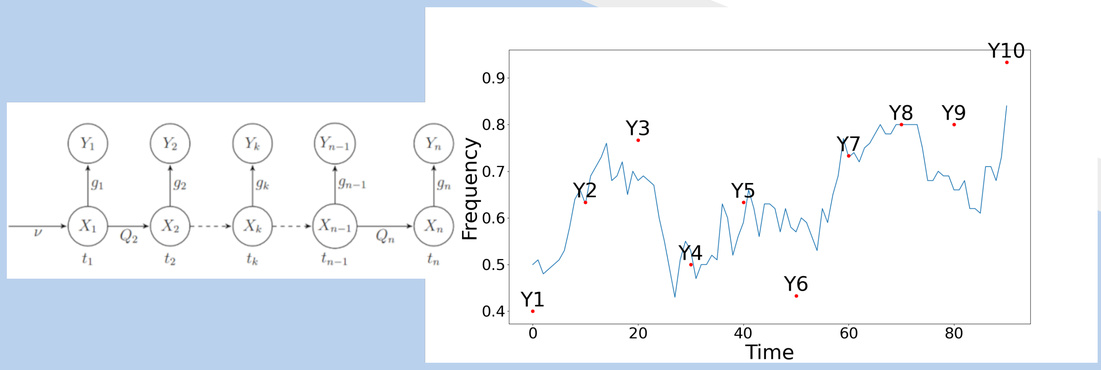

Figure 2: X_k Population frequency at time k, Y_k Sampling frequency at time k, Q_k Transition matrix from k-1 to k

The Hidden Markov Model (HMM) facilitates the rapid computation of the probability $P(Y_1, ..., Y_n|Q_1, ..., Q_n) = P(Y_1, ..., Y_n|N, s) = \ell(\mathbf{Y}; N, s)$ Here, $\mathbf{Y} = Y_1, ..., Y_n$ represents the observed sequence of data, and ℓ denotes the likelihood function. In this context, N stands for the population size and s for the selection coefficient.

Transition models

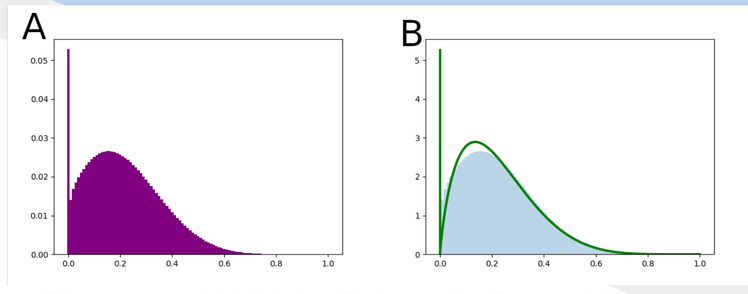

Figure 3: (A) Wright-Fisher, (B) Beta with spikes

In HMM, different models can be used to represent the transition probabilities between hidden states. The Wright-Fisher model is designed to closely mimic true evolutionary paths, but its complexity makes it difficult to integrate directly into HMM frameworks. Alternatively, approximations of the Wright-Fisher model with simpler distributions such as Gaussian and Beta are often used. However, it has been shown that the Beta with spikes distribution is the most accurate approximation of the Wright-Fisher model (2).

Results

When $s_1, ..., s_L$ are known, the optimization problem for maximizing $\ell(\mathbf{Y_1}, \mathbf{Y_2}, ..., \mathbf{Y_L}; N, s_1, s_2, ..., s_L)$ reduces to a one-dimensional task.

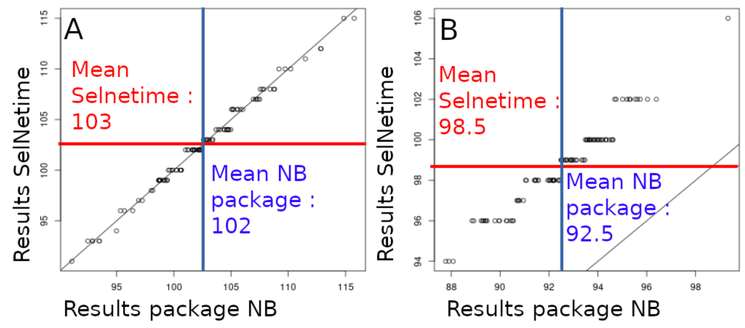

Figure 4: Comparison NB package, A dt = 1 B dt = 10

If N is known $s_k : \ell(Y_k; N, s_k), \forall k \in [1, L]$ can be optimized

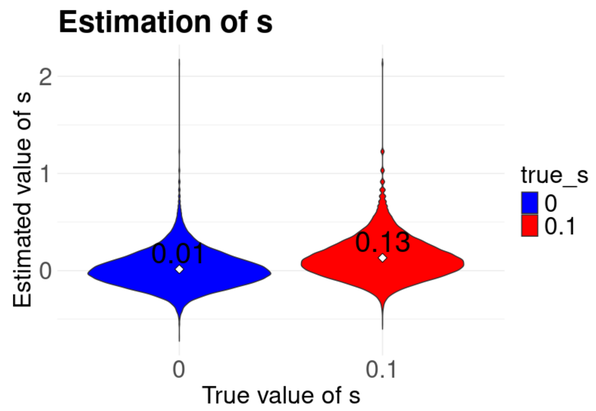

Figure 5: Error on the estimation of the selection parameter

IIn continuity with the work of Cyriel Paris (2), who had already shown the possibility of estimating s if N is known, our work confirms these results, but we've also shown that it is possible to estimate N if $s_1, ..., s_L$ are known.

Discussion

We've seen that increasing the time between each sampling point introduces biases in the estimation of the effective population size. Longer time intervals can lead to reduced genetic diversity and increased genetic drift, which can underestimate N.
Exploring all parameter combinations for optimisation is impractical. We plan to use a stochastic sampling algorithm that efficiently navigates high-dimensional spaces by iteratively sampling and updating parameters.
Logiciel simple d'utilisation : nom et lien

References

(1) Bollback, Jonathan P et al. "Estimation of 2Nes from temporal allele frequency data." Genetics (2008)
(2) Paris et al. "Inference of Selection from Genetic Time Series Using Various Parametric Approximations to the Wright-Fisher Model." G3 2019