Detection of loci under selection from temporal population genomic data through ABC random forest

Vitor Pavinato, Jean-Michel Marin, Miguel Navascués

Detection of loci under selection from temporal population genomic data through ABC random forest

Pavinato VAC[1,2]; De Mita S[3]; Marin JM[2]; Navascués M[1]

[1]UMR CBGP, INRA; [2]UMR IMAG, Université de Montpellier; [3]UMR IAM, INRA

POST-DOC Fellowship Program: InterLabEx Program

Key words: population genetics; adaptation; time-series, Approximation Bayesian Computation; machine learning; individual-based simulation

Abstract: Recent theoretical works have shown that the interaction between the signal of demography and selection can lead to bias in the inference of population size and the false identification of adaptive loci in genome scans. The joint estimation is a necessity, however not yet fully implemented. We propose the use of ABC Random-Forests to implement the joint inference in temporal population genomics datasets. Preliminary results showed that the method permits the joint inference of demography and selection, allowing distinguish the true demography (census size) and genetic drift (effective population size), as well the estimation of the genetic load (selection).

DESCRIPTION

Traditional population genetic studies use genotypic or allelic frequency data obtained from several populations sampled at the same time point. However, temporal population genetics data offers a more powerful way to study complex dynamics, since we can follow the allele frequency changes through time in the population.

Disentangling the effects of selection and demography is a long-standing difficulty in population genetics. Theoretical works show that selection affecting linked sites may bias inference of demography (Ewing and Jensen 2016; Schrider *et al.* 2016) and vice-versa. In this way, the ability to jointly make inference about both processes is a necessity not yet fully implemented. One proposed way to co-estimate neutral and selective parameters is with the use of ABC (Li *et al.* 2012); however, the traditional ABC is computationally expensive. The introduction of random forests in ABC, however, reduces the computational burden, making it possible to study complex dynamics as the presence of linked selection (Pudlo *et al.* 2016, Raynal *et al.* 2017).

We propose the use of ABC-RF to co-estimate demography and selection in temporal population genomics data.

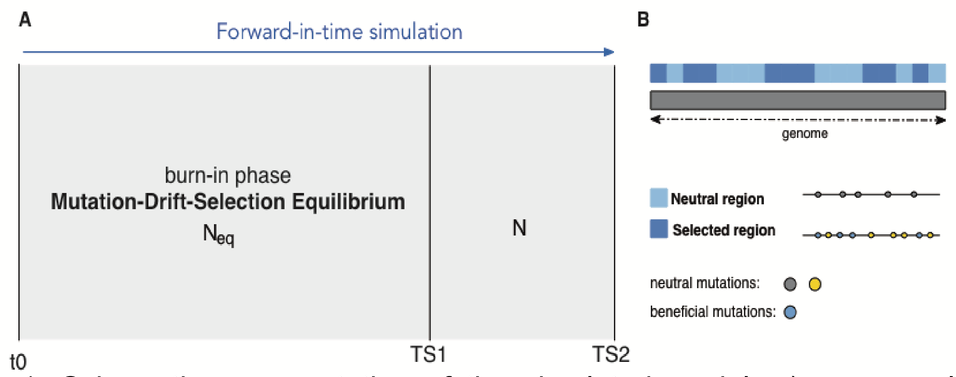

Figure 1. Schematic representation of the simulated model: a) one population sampled at two time points and; b) the genomic architecture that it a combination of neutral and beneficial mutation defined by the prior.

The simulation part of our ABC framework was conducted with SLiM version 3.1 (Haller and Messer 2017). We simulated a simple scenario where a population of size *N* was sampled twice at the begin and at the end of a time interval. The proportion of the genome, the number of mutations under selection, and the mean of the gamma distribution that defines the distribution of fitness effects were sampled from prior probability density distribution.

We ran ~50.000 simulations and calculated summary statistics to construct the reference table that was used to row the random-forests.

RESULTS

To test if the grown random-forest can recovers the expected population size *N* and the genetic load (selection parameter) we ran an additional 100 simulations with and without selection and used them as pseudo-observed data-set (POD).

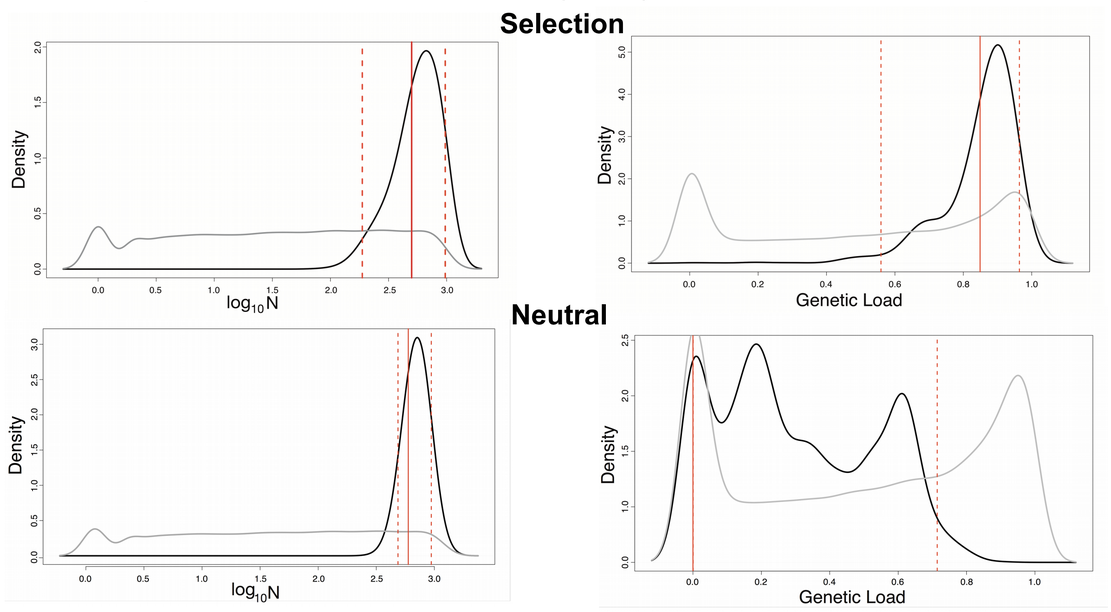

Figure 2. The prior (grey) and posterior (black) distributions of the population size *N* and the genetic load of one POD. The horizontal red lines are the true value, and the dashed lines represent the credible intervals. Top: the model with selection and, down: the model without selection.

We compared the ABC-RF estimated population size *N* with the moment-based estimated effective population size N_e. The moment-based estimator is based on the averaged allele frequency differences between the time-points and is affected by the neutral (drift) and non-neutral dynamics (selection).

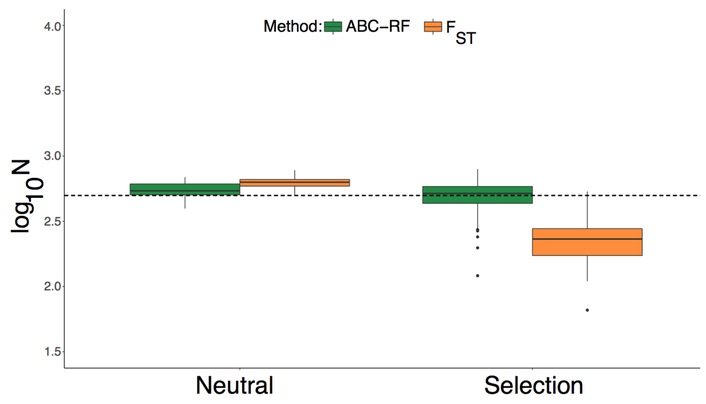

Figure 3. The comparison between ABC-RF and moment-based methods to estimate the population size. Dashed line represent the true value.

Ewing GB, Jensen JD (2016). Mol Ecol 25: 135–141.
Li J, Li H, Jakobsson M, Li S, Sjödin P, Lascoux M (2012). Mol Ecol 21: 28–44.
Pudlo P, Marin J-M, Estoup A, Cornuet JM, Gautier M, Robert CP (2016). Bioinformatics 32: 859–866.
Raynal L, Marin J-M, Pudlo P, Ribatet M, Robert CP, Estoup A (2017). arXiv: 1605.05537.
Schrider DR, Shanku AG, Kern AD (2016). Genetics 204: 1207–1223.