# Joint inference of demography and selection from temporal population genomic data via approximate Bayesian computation

Vitor Antonio Correa Pavinato, Stéphane de Mita, Jean-Michel Marin, Miguel Navascués

# Joint inference of demography and selection from temporal population genomic data via ABC

Vitor A.C. Pavinato  Stéphane De Mita  Jean-Michel Marin  Miguel de Navascués

INRAE, Uppsala universitet, Université de Montpellier miguel.navascues@inrae.fr

## Common assumptions in population genetic inference

Most classical population genetic inference methods assume that genome wide genetic diversity is mostly influenced by neutral processes (commonly named as "demography") while selective processes affecting few isolated loci. Thus, changes in allele frequencies through time for a large number of loci can be used to infer the amount of drift (i.e. the effective population size, $N_e$) and the presence of some selection is assumed to have a negligible impact on that estimate. Loci under selection can then be identified as they will present allele frequency changes larger or more often in the same direction, than expected under pure drift.

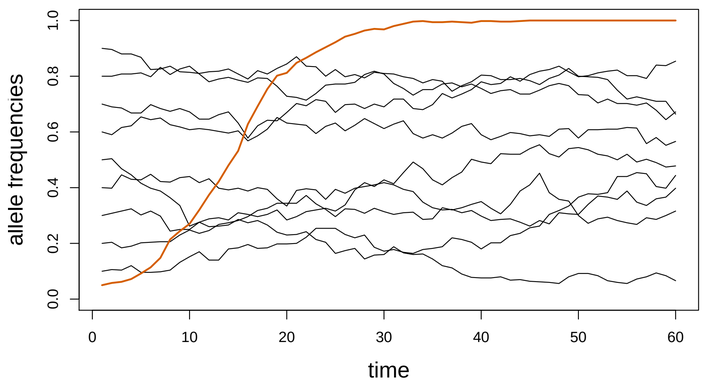


Figure: **A naive view of population genetics.** Simulation of allele frequencies change through time in nine neutral loci (black) and one adaptive locus (red) due to drift ($N_e = 500$) and selection ($s = 0.4$).

## Reality is more complex

In some cases these assumptions might not hold. If the action of selection is widespread along the genome or recurrent in time, the high proportion of loci affected by selection and hitch-hiking could significantly bias the estimates of effective population size. Removing outlier loci would not solve this problem because only loci with large effects will be detected but many other loci under weaker selection (e.g. selection on polygenic characters) could be present.

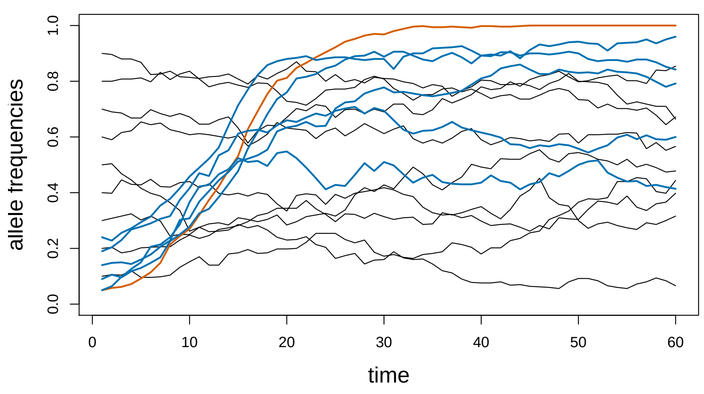


Figure: **A proportion of neutral loci are afected by linked selection.** Simulation of allele frequencies change through time in nine neutral loci (black), one adaptive locus (red) and five hitch-hiking neutral loci (blue) due to drift ($N_e = 500$), selection ($s = 0.4$) and recombination ($0.005 \leq r \leq 0.1$).

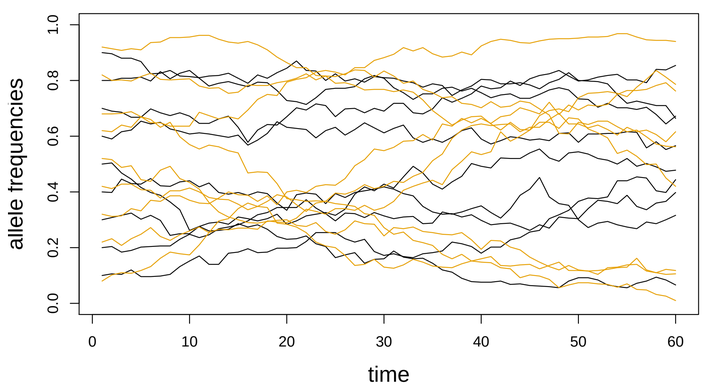


Figure: **Loci under weak selection are difficult to distinguish from neutral loci.** Simulation of allele frequencies change through time in nine neutral loci (black) and nine adaptive loci (orange) due to drift ($N_e = 500$) and selection ($0.03 < s < 0.13$).

## Our proposal: joint inference using ABC

We propose to make population genetic inference using models that include both the neutral and adaptive processes on the genome scale. This will allow to estimate demographic and selection parameters taking into account their interaction. Because these models are difficult to address under a likelihood framework we recourse to approximate Bayesian computation *via* Random Forest (ABC-RF). ABC-RF uses simulations to generate a training data set from which RF can learn and make inferences from real data.

## Our model

We applied this approach to the case of a single population of size $N$. Advantageous mutations arrive to the population at rate $\mu_b$ with selection coefficients ($s$) taken from a gamma distribution. We simulate this model with SLiM and at each generation the effective population size ($N_e$) is calculated from the variance of reproductive success of individuals. We also calculate genetic load ($L$) and proportion of polymorphisms with $Ns > 1$ ($P$). Samples of individuals are taken at two different generations, separated by a period of time $\tau$.

## Results: Estimation of $N_e$

The performance of the method was assessed using the simulations from the training data. For each of the the calculated $N_e$ during the simulation (true $N_e$) was compared to its estimate (out-of-bag, OOB, prediction from ABC-RF). An estimate of $N_e$ from temporal $F_{ST}$ was also calculated as comparison with a method that assumes all loci to be neutral.

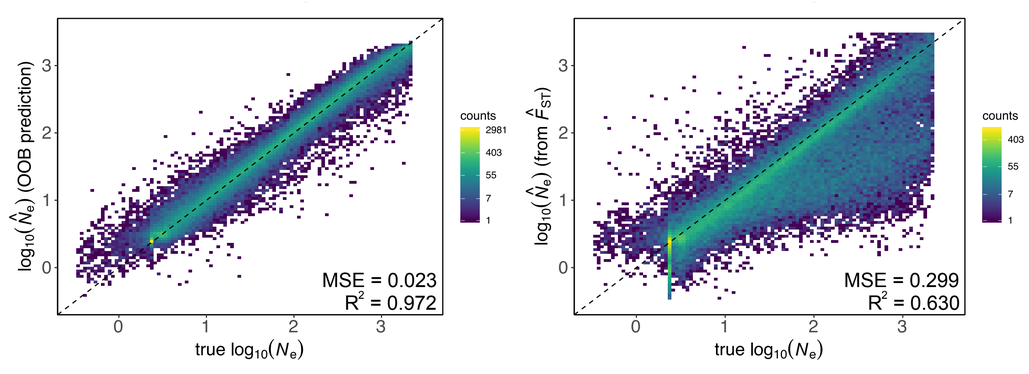


Figure: **Performance of the $N_e$ estimation.** Left: Estimates from ABC-RF. Right: Estimates from temporal $F_{ST}$ (assume no selection).

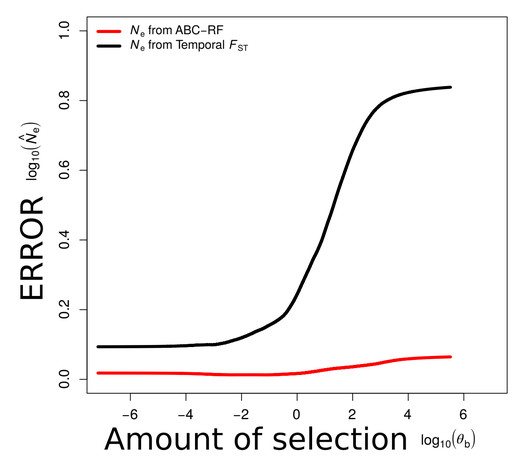


Figure: **Good $N_e$ estimation despite pervasive selection.** Error in $N_e$ estimate in function of the presence of selection (measured as $\theta_b = 4N_e\mu_b$). Estimates from $F_{ST}$ show a dramatic increase in error when selection is frequent in the population but ABC-RF estimates are only weakly affected.

## Results: Estimation of selection

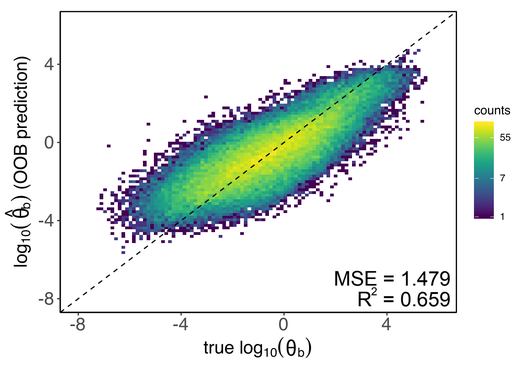


Figure: **Quantities measuring selection are a bit harder to estimate.** We can also obtain estimates for genetic load and proportion of polymorphisms under selection (see preprint).

## Conclusions

- Including both neutral and adaptive processes in the model allows:
  - Good estimate of demography in presence of pervasive selection
  - Estimate of quantities describing the action of selection based exclusively on polymorphism data.
- Further work needed to evaluate more complex models

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

**Acknowledgements**

This project has received funding from the LabEx AGRO (convention ANR-10-LABX-0001-01), CEMEB (convention ANR-10-LABX-0004) and NUMEV (convention ANR-10-LABX-20) through the AAP Inter-LabEx (ABCSelection). This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 791695 (TimeAdapt). S. De Mita was funded by INRAE (*Projet Innovant EFPA*).