Detecting selection from genomic time series: the Beta with Spikes approximation

Cyriel Paris, Bertrand Servin, Simon Boitard

Detecting selection from genomic time series : the Beta with Spikes approximation

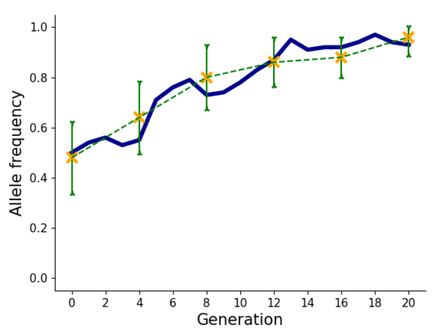
Cyriel PARIS, Bertrand SERVIN, Simon BOITARD
INRA Toulouse, GenPhySE, France

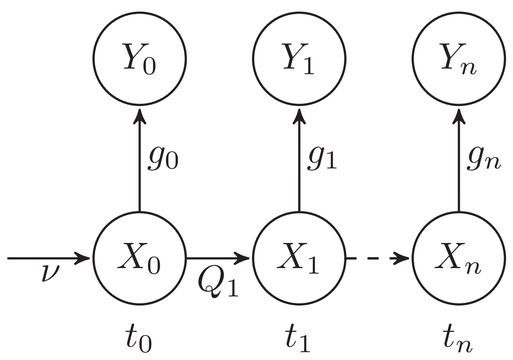
Abstract

Detecting genomic regions under selection is an important objective of population genetics. Typical analyses for this goal are based on exploiting genetic diversity patterns in present time data but rapid advances in DNA sequencing have increased the availability of time series genomic data. A common approach to analyze such data is to **model the temporal evolution of an allele frequency as a Markov chain**. Based on this principle, several methods have been proposed to infer selection intensity. One of their differences lies in how they model the **transition probabilities** of the Markov chain. Using the **Wright-Fisher model** is a natural choice but its **computational cost** is **prohibitive** for large population sizes so approximations to this model based on parametric distributions have been proposed. Here, **we compared** the performance of **some** of these **approximations** with respect to their power to detect selection and estimation of the selection coefficient. We developped a new generic Hidden Markov Model likelihood calculator and applied it on genetic time series simulated under various evolutionary scenarios. The **Beta-with-Spikes** approximation, which combines discrete fixation probabilities with a continuous Beta distribution, was found to perform **consistently better than the others**. This distribution provides an almost perfect fit to the Wright-Fisher model in terms of selection inference, for a computational cost that does not increase with population size. Based on this model, we show that the **sampling period** considered strongly determines the selection intensities that can be detected or properly estimated.

Theoretical framework [1]

- Population **allele frequency trajectory** at a locus impacted by selection: increases for beneficial alleles.
- Observed only at **a few dates** in a **sample of the population**.





- **Hidden Markov Model (HMM)**: population allele frequency is a hidden Markov chain (X_{t_i}), observations (Y_{t_i}) are sample allele frequencies.
- Data likelihood efficiently computed using the Forward algorithm.
- Transition model (Q_i) ?

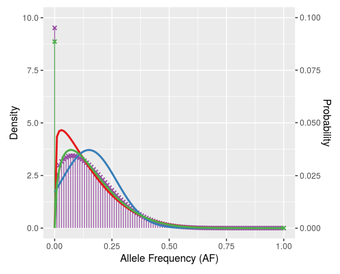
We developped (in python) a generic likelihood calculator and **compared selection inference for different scenarios and transition models**.

A reference model : Wright Fisher [2]

- Bi-allelic locus (A_0/A_1), X_t frequency of allele A_1 at time t.
- Random mating : $X_{t+1}|X_t \sim \frac{1}{N_e}\mathcal{B}(N_e, f(X_t))$
Genotype	A_1A_1	A_1A_0	A_0A_0
Fitness	$1+s$	$1+sh$	1
- Fitness function: $f(x) = \frac{(1+s)x^2+(1+sh)x(1-x)}{(1+s)x^2+2(1+sh)x(1-x)+(1-x)^2}$
- Matrix Q_i size $N_e \times N_e \rightarrow$ **numerically intractable for large** N_e.

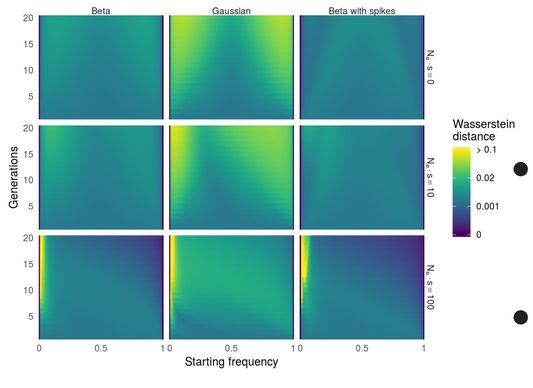
Alternative models : moment fitting

- Choose a **parametric distribution** (Beta, Gaussian, Beta-with-Spikes) [3, 4, 5]
- Approximate Wright Fisher (WF) moments with a recursion using Taylor expansions. [3, 4, 5]



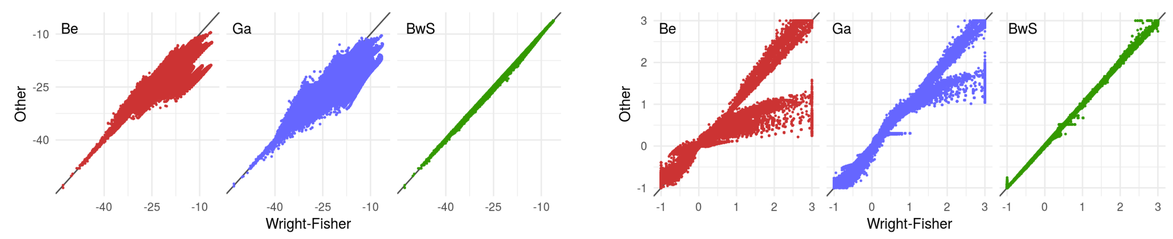
- **Fit moments** of the parametric distribution **to those of the WF**.
- Compute data likelihood and infer selection using this distribution.

Transition approximation



- Transition matrix (Q_i) of each parametric distribution is compared to that of the WF model using the Wasserstein distance.
- **Accuracy** of all approx. **decreases for long times** and **starting frequencies close to 0 or 1**.
- **Beta-with-Spikes** model is overall a **better approximation of WF** transitions.

Likelihood approximation ($N_e = 100$)

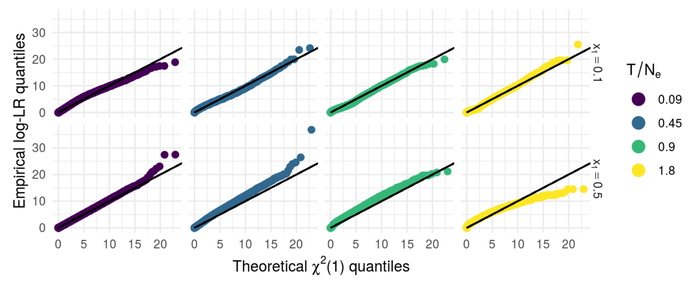


- Likelihood profiles evaluated with all transition models, for datasets simulated under WF with different s and initial frequency x_0.
- The **Beta-with-Spikes closely fits the WF** for e.g. the **likelihood at** $s = 0$ (left) or the **Maximum Likelihood Estimator** $\hat{s}$ (right).

Beta-with-Spikes : statistical properties

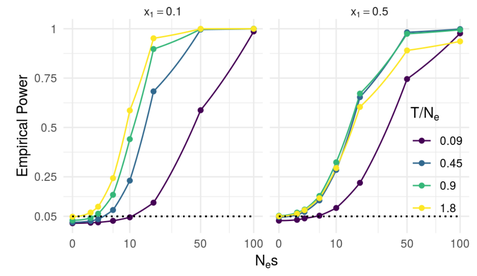
Quality of inference studied based on simulated datasets with 10 sampling dates, for different N_e (100, 1000, 10000), x_0 and evolution times T.

- **Selection detection:** the null hypothesis: "$s = 0$" is tested using a Likelihood Ratio (LR) test.

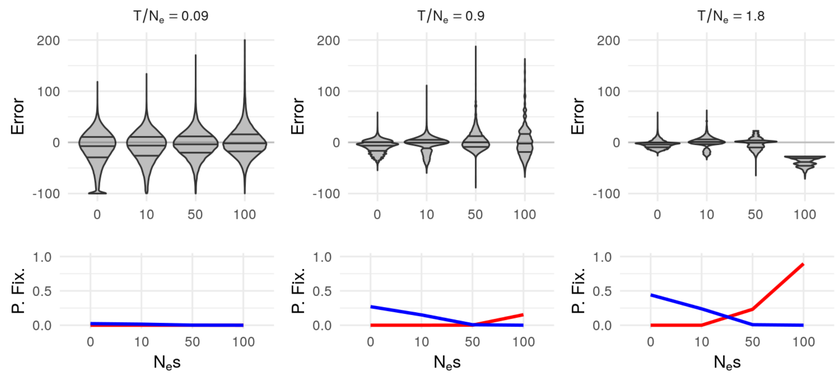


Calibration: the LR distribution under $s = 0$ follows a $\chi^2(1)$ $\rightarrow$ Null hypothesis rejected based on $\chi^2(1)$ quantiles.

Detection power increases with $N_e s$ and (up to a certain point) with T/N_e.



- **Estimation of selection intensity (s):**
  - — **MLE generally unbiased**, decreased variance for large T/N_e (top).
  - — Negative bias for large T/N_e and $N_e s$, due to a high number of trajectories fixing A_1 between the two first sampling dates (bottom).



References

[1] Olivier Cappé, *Inference in Hidden Markov Models*, Springer
[2] Warren J. Ewens, *Mathematical Population Genetics*, Springer
[3] Lacerda & Seoighe, Genetics, Vol 198, 1237-1250, November 2014
[4] Terhorst *et al.*, PLOS genetics, 11(4): e1005069
[5] Tataru *et al.* Genetics, Vol 201, 1133-1141, November 2015