Comparison of differents models infering selection from genomic time series data

Cyriel Paris, Simon Boitard, Bertrand Servin

Comparison of differents models infering selection from genomic time series data

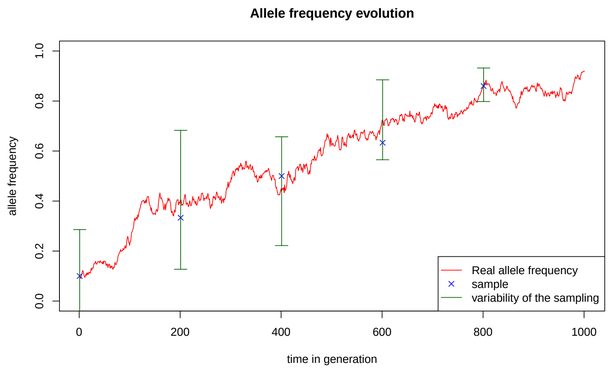
Cyriel PARIS, Bertrand SERVIN, Simon BOITARD
INRA Toulouse, GenPhySE, France

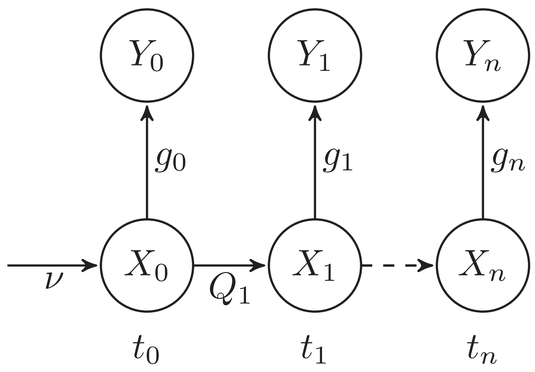
Introduction

Natural or artificial **selection** have a lot of impact on a population genetic pool. When a **beneficial mutation** occurs in a group of individuals, these carrying this new gene are more adapted than the others to their environment and so have a better chance to reproduce and spread this new mutation. Given genetic samples, there exists a lot of **method detecting the genome regions under selection**. These methods use **present data samples**. However, new genotyping techniques give genomic samples through time (From few decades to centuries). This new kind of data needs a new methodology to exploit this information. There already exists few recent methods using time series data to detect selection mainly concerning ancient DNA. The **objective** of my PhD is to design an algorithm detecting loci under selection given **genomic time series**. The purpose of this poster is to show differences between few theoretical models concerning selection detection and inference.

Theoretical framework [1]

- Selection impacts the part of populations carrying the selected allele (AF).
- Only some times are observed
- Each observation has a variability





- Hidden Markov Model takes into account these conditions
- AF is a hidden markov process
- Observations depends only of the current AF
- Efficient Forward / Backward algorithm to get likelihood of data

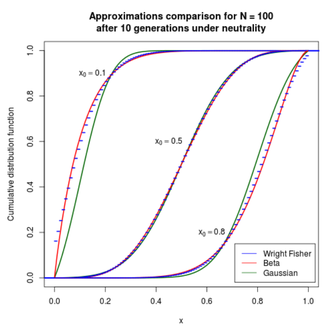
We developped (in python) a generic Likelihood calculator to compare selection detection / estimation under various scenarios and AF transition models.

A reference model : Wright Fisher [2]

- Single Nucleotide Polymorphism (SNP)
- X_t : derived allele frequency at time t
- Random mating : $X_{t+1}|X_t \sim \frac{1}{N}\mathcal{B}(N, f(X_t))$
Genotype	A_1A_1	A_1A_0	A_0A_0
Fitness	$1+s$	$1+sh$	1
- fitness function: $f(x) = \frac{(1+s)x^2+(1+sh)x(1-x)}{(1+s)x^2+2(1+sh)x(1-x)+(1-x)^2}$
- Not numerically tractable when N_e is large

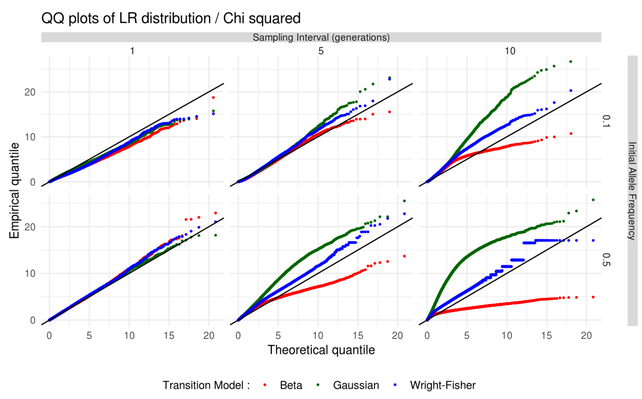
Alternative models : moment fitting

- Choose a parametrised distribution (beta, gaussian) [3, 4, 5]
- Approximate Wright Fisher process moments with a recursion derived from a Taylor expansion. [3, 4, 5]



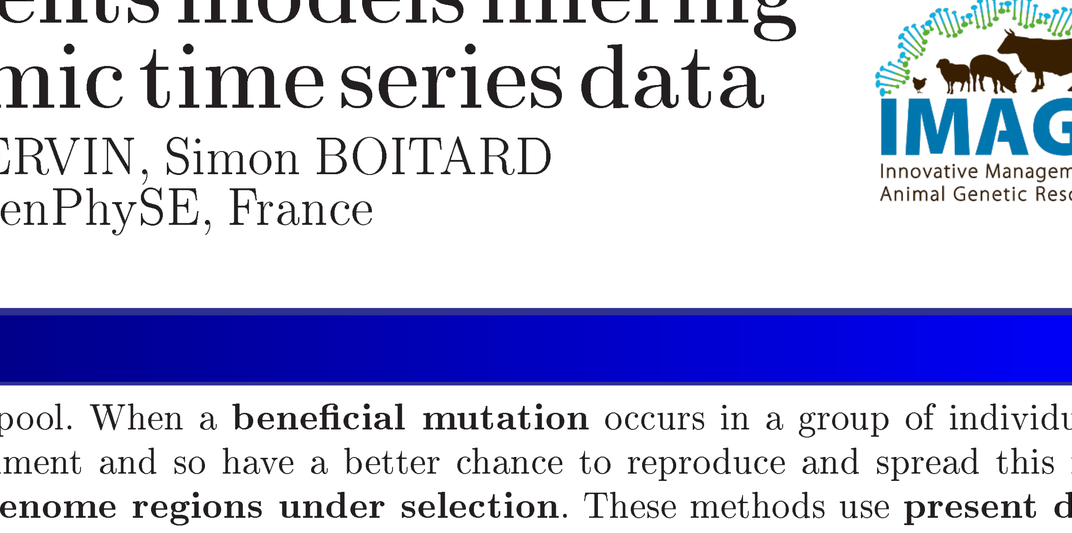
- Fit moments of this distribution with the moment approximation
- Compute likelihood of observations under the choosen model
- Use Likelihood Ratio statistic (LR) to detect selection via the Likelihood Ratio Test (LRT)

Calibration under neutrality

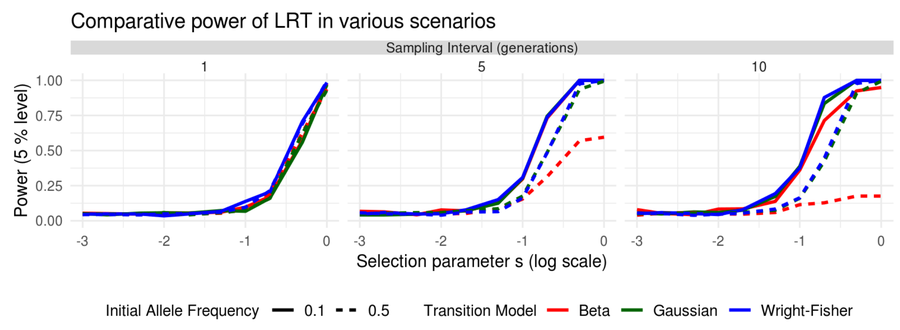


- $LR \sim \chi^2$ only when fixations events are not likely.
- Only empirical null distributions are used to get statistical power

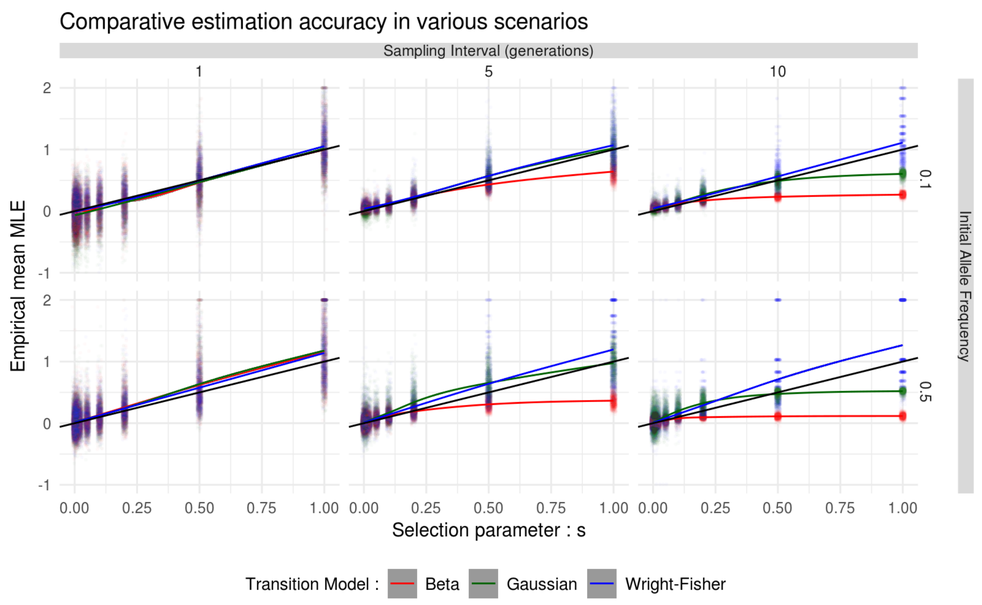
Detecting selection (on simulations)



- Time range has to be long enough to see variations
- Time range has to be short enough that allele is not fixed in population



- Statistical power increase as selection parameter is getting bigger
- Starting from $x_0 = 0.5$, selection is harder to detect if sampling time interval increase



- The Maximum likelihood estimator (MLE) is consistent with WF
- Approximations underestimate high selection parameter when sampling time interval increase.
- The estimator becomes less accurate (higher variance than low selection case) while selection increase

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