

Comparison of differents models infering selection from genomic time series data

Cyriel Paris, Simon Boitard, Bertrand Servin

▶ To cite this version:

Cyriel Paris, Simon Boitard, Bertrand Servin. Comparison of differents models infering selection from genomic time series data. 2. Joint Congress on Evolutionary Biology (EVOLUTION 2018), Aug 2018, Montpellier, France. 2018. hal-02788098

HAL Id: hal-02788098 https://hal.inrae.fr/hal-02788098v1

Submitted on 5 Jun2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



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]

 Y_1

 g_1

lele frequency evolution

DETECTING SELECTION (ON SIMULATIONS)

Impact of sampling time and selection parameter on allele frequency evolution

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

 Y_0



- Hidden Markov Model takes into account these conditions
- AF is a hidden markov process
 Observations depends only of the
- Observations depends only of the current AF

 $\overbrace{t_0}^{\Lambda_0} \overbrace{Q_1}^{\Lambda_1} \overbrace{t_1}^{\Lambda_1} \overbrace{t_n}^{\Lambda_n} \bullet \text{ Efficient Forward / Backward al-} \\ t_0 \quad t_1 \quad t_n \quad \text{gorithm to get likelihood of data} \\ \text{We developped (in python) a generic Likelihood calculator to compare} \\ \text{selection detection / estimation under various scenarios and AF transition} \\ \text{models.}$

A REFERENCE MODEL : WRIGHT FISHER [2]

 I_n

 $|g_n|$



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



- 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 Fitness1+s1+sh1

- 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]



Selection parameter s (log scale)

itial Allele Frequency — 0.1 - 0.5 Transition Model — Beta — Gaussian — Wright-Fisher

- 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



Comparative estimation accuracy in various scenarios

- The Maximum likelihood estimator (MLE) is consistent with WF
- 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.

pproximations comparison for N = 100 after 10 generations under neutrality

> Wright Fisher Beta

 Only empirical null distributions are used to get statistical power

- 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

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