SelNeTime: a python package inferring effective population size and selection intensity from genomic time series data

Mathieu Uhl[1,3], Miguel Navascués[1], Simon Boitard[1], Bertrand Servin[2]

[1]CBGP, Université de Montpellier, CIRAD, INRAE, IRD, Institut Agro Montpellier, Montpellier, France
[2]GenPhySE, Université de Toulouse, INRAE, INPT, Castanet-Tolosan, France
[3]CEFE, CNRS, Montpellier, France

2024-11-06

Summary

Genomic samples collected from a single population over several generations provide direct access to the genetic diversity changes occurring within a specific time period. This provides information about both demographic and adaptive processes acting on the population during that period. A common approach to analyze such data is to model observed allele counts in finite samples using a Hidden Markov model (HMM) where hidden states are true allele frequencies over time (i.e. a trajectory). The HMM framework allows to compute the full likelihood of the data, while accounting both for the stochastic evolution of population allele frequencies along time and for the noise arising from sampling a limited number of individuals at possibly spread out generations. Several such HMM methods have been proposed so far, differing mainly in the way they model the transition probabilities of the Markov chain. Following Paris et al. (2019a), we consider here the Beta with Spikes approximation, which avoids the computational issues associated to the Wright-Fisher model while still including fixation probabilities, in contrast to other standard approximations of this model like the Gaussian or Beta distributions. To facilitate the analysis and exploitation of genomic time series data, we present an improved version of Paris et al. (2019a) 's approach, denoted SelNeTime, whose computation time is drastically reduced and which accurately estimates effective population size in addition to the selection intensity at each locus. This method is implemented in a user friendly python package, which also allows to easily simulate genomic time series data under a user-defined evolutionary model and sampling design.

Introduction

One limitation of most population genetic analyses is to use data from a single time point, the present. The genetic diversity observed in this single sample is very informative about past population history, but it integrates the effect of many evolutionary processes acting on a large time-scale, which may be hard to clearly separate. For instance, selection constraints acting on a given locus often fluctuate along time as a result of environmental changes, and such complex and changing processes are difficult to reconstruct from the observation of the resulting genetic diversity at a single recent time point. In contrast, the study of samples collected at several different generations allows to directly measure changes in the genetic diversity of populations, informing about the evolutionary processes acting on that specific time period. Such genomic time series arise in different contexts and time scales, such as the monitoring of wild or domestic populations (Boitard et al., 2023), experimental evolution (Long et al., 2015) or ancient DNA (Fages Antoine et al., 2019) studies.

A common approach to analyze genomic time series data is to model observed allele frequencies using a Hidden Markov model (HMM); this approach allows computing the full likelihood of the data, while accounting both for the stochastic evolution of population allele frequencies along time, under the combined effects of demography and selection, and for the noise arising from sampling a limited number of individuals at different generations (Bollback et al., 2008). Several such HMM methods have been proposed so far, differing mainly in the way they model the transition probabilities of the Markov chain. Indeed, while the Wright-Fisher (WF) model appears as a natural choice to model these transitions, it leads to transition matrices whose dimension is equal to the population size so it is limited in practice to small populations. Among the various approximations (e.g. Gaussian, Beta) that have been proposed to overcome this issue, the Beta with Spikes (BwS) distribution (Tataru et al., 2015) was pointed out by Paris et al. (2019a) as particularly promising, because it combines a continuous (Beta) density with discrete fixation probabilities in 0 and 1. These features allow accurate estimations for a large range of demographic and selective parameters (Paris et al., 2019a), while previous methods were either impossible to run in large populations or inaccurate for some parts of the parameter space (small populations, strong selection).

However, the current implementation of this method, compareHMM (Paris et al., 2019b), considers a fixed effective population size provided by the user, assuming that this parameter is known or has been previously estimated from another software. Besides, it was initially implemented as a test tool comparing different approximations of the Wright-Fisher model in the context of genomic time series analyses, not as a real package to be used by a large community. In particular, it was not optimized to analyze very large data sets such as those typically found in the genomics era.

With the objective of improving on these several limitations, the SelNeTime python package (Uhl et al., 2024) provides a user friendly and computationally efficient implementation of Paris et al. (2019a)'s method, allowing to estimate effective population size and selection from genomic time series and to simulate such data in order to easily test the expected performance of the approach under custom sampling designs. This package is available on

the python package index https://pypi.org/project/selnetime/ and on the project git repository https://forgemia.inra.fr/simon.boitard/snt.

Important features

Estimation of effective population size

Given an observed temporal sequence of sampled allele counts $\mathbf{Y}_i$ at a given locus i, the HMM approach described in Paris et al. (2019a) can evaluate the (exact) likelihood $P(\mathbf{Y}_i|N, s_i)$ for any values of the effective population size N and of the selection intensity s_i. As indicated in the notations, s_i is a locus specific parameter, because selective contraints typically differ between loci. In contrast, N reflects the demography of the population so a single value applies to the whole genome. Considering p loci assumed to be independent (which almost holds in practice if these loci are taken sufficiently far away from each other) and neglecting in first approximation the effect of selection at each of these loci, SelNeTime computes the likelihood of N as

$$P(\mathbf{Y}_1, \dots, \mathbf{Y}_p|N) = \prod_{i=1}^{p} P(\mathbf{Y}_i|N, s_i = 0)$$

and finds the Maximum Likelihood Estimator (MLE) of N, i.e. the value of N maximizing the above formula.

Using computer simulations, we checked that this approach provides accurate estimates of N, as illustrated in Figure 1. As a comparison, we estimated N for the same data with the method of Hui & Burt (2015), implemented in the NB package (Hui, 2014), which is based on a continuous Beta approximation of allele frequency transitions. Interestingly, both methods performed equally well for short sampling time intervals ($dt = 1$) but the accuracy of NB was reduced for larger time intervals ($dt = 10$); this is consistent with the higher probability of fixation events, which are not accounted in the method of Hui & Burt (2015), when dt increases, illustrating the advantage of the BwS approximation.

Estimation of selection intensity at a given locus

Once the effective population size N has been estimated, SelNeTime estimates the selection intensity at each locus i using the same approach as Paris et al. (2019a), i.e. optimizing $P(\mathbf{Y}_i|N, s_i)$ over different values of s_i in order to find the MLE of this parameter. This leads to unbiased estimations of the selection intensity (Figure 2), as already explored in details by Paris et al. (2019a) for a large range of parameter values and various sampling designs.

Improved computation time

Another important improvement of SelNeTime compared to the inital implementation of the BwS approach in compareHMM (Paris et al., 2019b) concerns the computation time, which was reduced using two main actions. First, the continuous density of the distribution was discretized when building the transition matrices (with a controlled bin width), allowing this

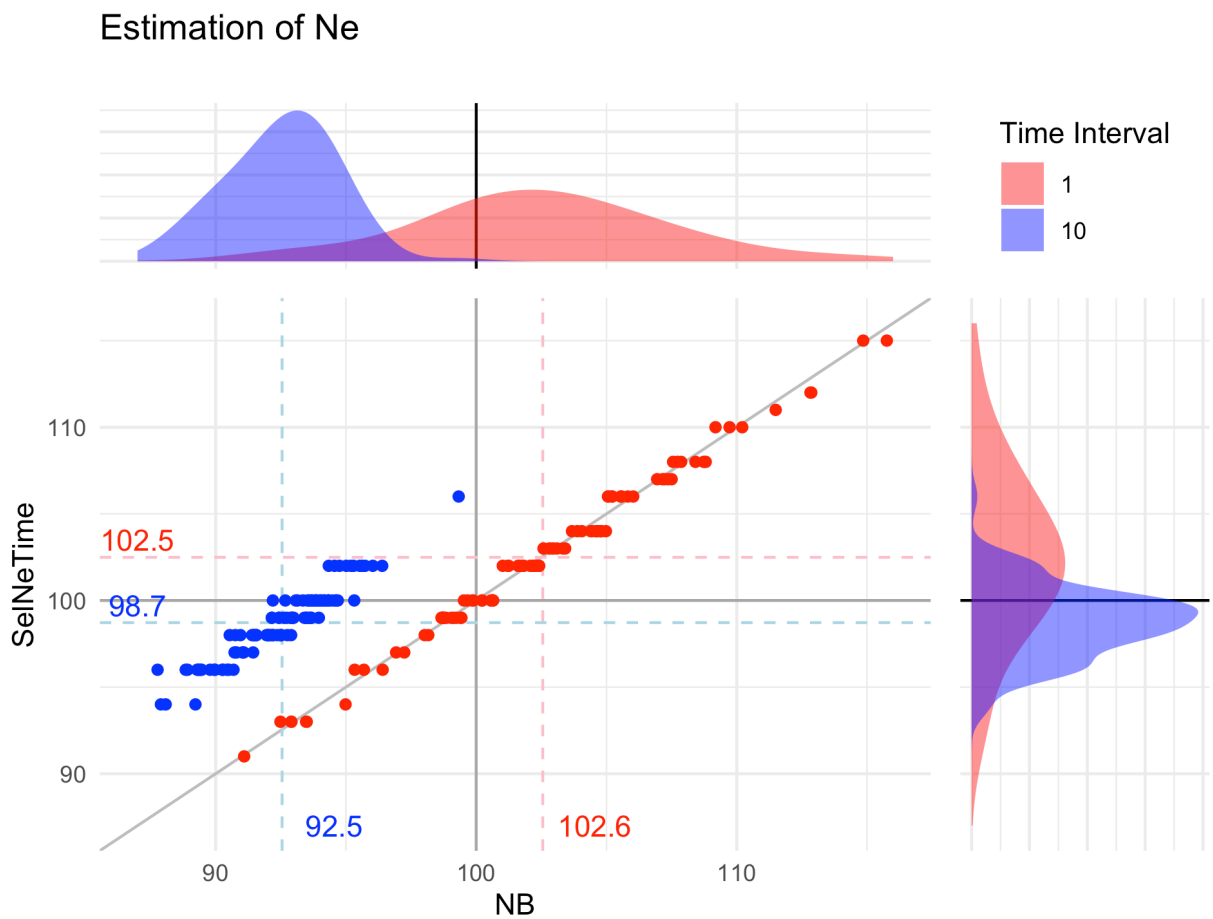

Figure 1: Estimation of the effective population size for 100 datasets, using SelNeTime or NB (Hui, 2014). Each dataset includes 1000 loci and 10 sampling times $t = 1 \ldots 10$ (red) or $t = 1, 10, \ldots, 100$ (blue). Allele frequency trajectories at each locus were simulated under the Wright-Fisher model with $N = 100$, $s = 0$ and an initial allele frequency $\pi_0 = 0.5$. 30 alleles were observed at each sampled generation.

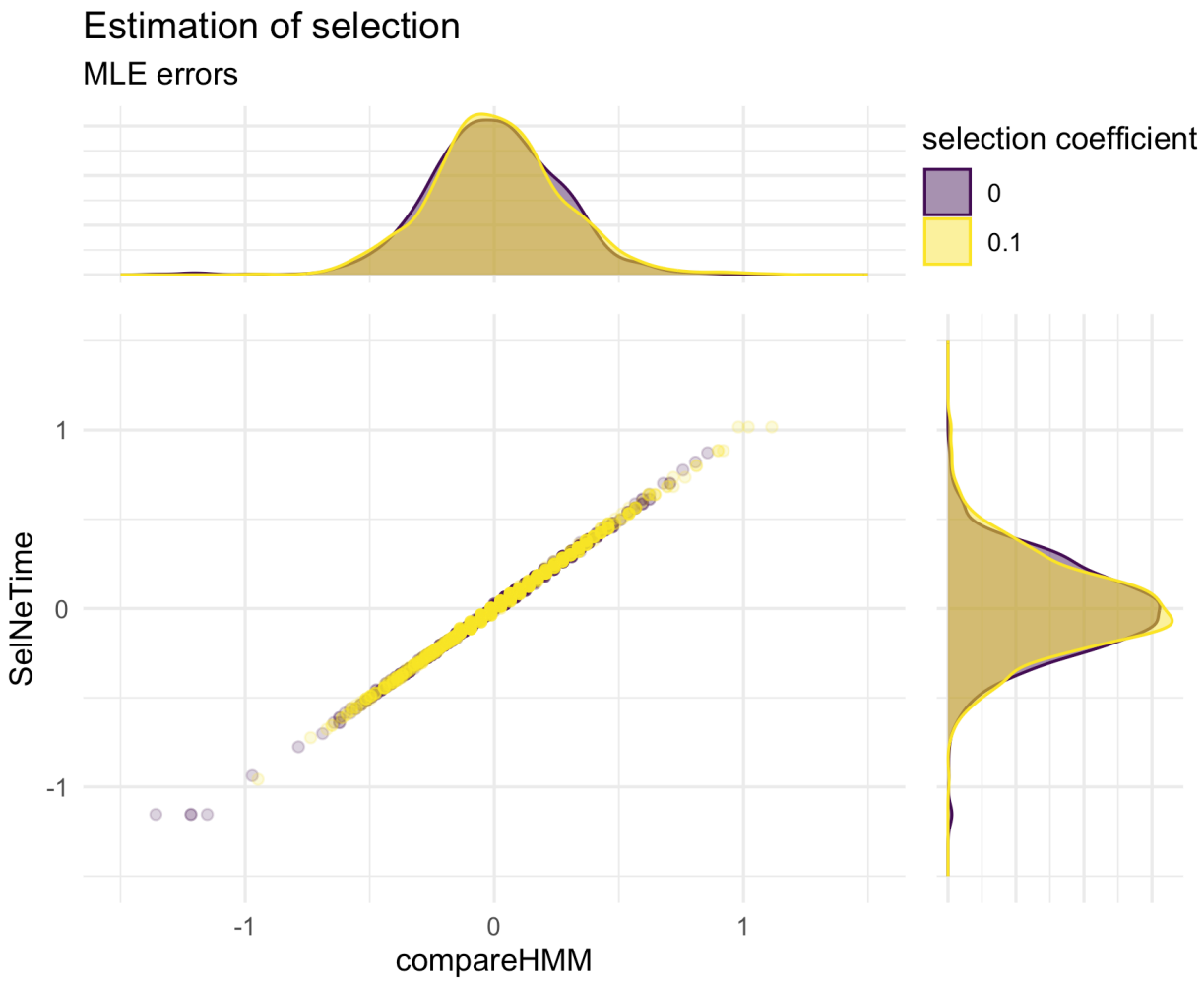

Figure 2: Estimation of the selection intensity for 1000 loci using SelNeTime. Allele frequency trajectories at each locus were simulated under the Wright-Fisher model with $N = 100$, $s = 0$ (purple) or $s = 0.1$ (yellow) and an initial allele frequency $\pi_0 = 0.5$. 30 alleles were observed at 10 sampling times $t = 1 \ldots 10$.

continuous part of the distribution to be analyzed exactly in the same way as the discrete probability mass points in 0 and 1. Second, matrix computations performed at each iteration of the HMM forward algorithm were vectorized, taking advantage of the Numpy library (Harris et al., 2020) to efficiently handle different loci or parameter values in a single operation.

As an illustration, Table 1 shows the computation time required to analyze a time series of 10 samples of 30 alleles each, with a sampling time interval of $dt = 10$ generations and a number of loci varying from 100 to 10000.

Table 1: Computation times of CompareHMM (Paris et al., 2019b) and SelNeTime

Nb. loci	compareHMM, estimation of s	SelNeTime, estimation of s	SelNeTime, estimation of N
100	39.27s	28.7s	6.12s
1000	360.01s	36.21s	14.23s
10000	3530.47s	96.95s	87.83s

For 100 loci, SelNeTime runs in approximately the same time as compareHMM, but the relative performance of SelNeTime improves quickly when the number of loci increases. Actually, the time to compute all transition matrices, which is the first step of both SelNeTime and compareHMM analyses, is independent of the number of loci (but it depends on the sampling design). This part of the analysis took approximately 28s for SelNeTime in this example, which was of the same order as compareHMM, but after this the time spent per locus was far lower with SelNeTime (approx. 0.007s) than with compareHMM (approx. 0.35s).

Comparison with the Wright-Fisher model

While SelNeTime computes by default HMM transitions using the BwS distribution, it might occasionally be useful to compute them based on the WF model, in particular to evaluate the impact of the BwS approximation on inference quality. Although this is numerically possible only for small population sizes (not more than a few hundreds), we included one option to do this in the program. For a true population size of $N = 100$, we could for instance check that using the BwS or the WF model had very limited impact on the inference of either effective population size or selection (Figure 3). A more detailed comparison between the estimations obtained by the BwS and WF distributions can be found in Paris et al. (2019a), but only for selection inference.

Simulation of datasets

As already observed in Paris et al. (2019a) and several other studies on the subject, the estimation accuracy of effective population size or selection depends on both the true evolutionary parameters of the population under study (N and s) and the sampling design of observed data (number and size of observed samples, time intervals between samples). In order to help researchers plan future experiments, or interprete the results obtained for

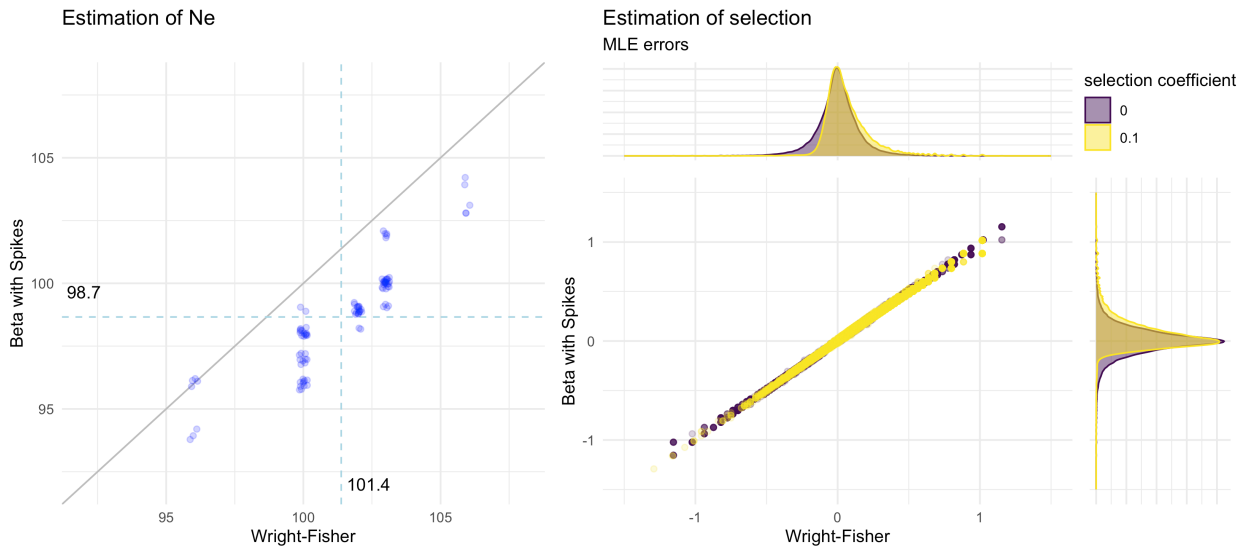

Figure 3: Comparing of SelNeTime estimations based on the Beta with Spikes and the Wright-Fisher model. Left: MLE of effective population size for 100 datasets of 1000 loci each. (B) MLE of s for 1000 loci. The two panels show the estimations obtained using the WF (x axis) or the BwS (y axis) model on the same data, which were simulated under the Wright-Fisher model with $N = 100$ $s = 0$ (Left and Right, purple) or $s = 0.1$ (Right, yellow) and initial allele frequencies $\pi_0 = 0.5$. 30 alleles were observed at 10 sampling times $t = 1, 10 \ldots 100$.

a specific dataset, SelNeTime includes a simulator of observed allele frequency trajectories. Simulations can easily be tunned using either a yaml file or command line options and output data at a format that can directly be analyzed with other SelNeTime commands. Notably, this simulator was used to generate the data underlying the figures and tables of this study.

Acknowledgments

This study has recieved the support of the Occitanie Regional Council's program 'Key challenge BiodivOc'. We are also grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing computing resources.

References

Boitard S, Liaubet L, Paris C, Fève K, Dehais P, Bouquet A, Riquet J, Mercat M-J (2023) Whole-genome sequencing of cryopreserved resources from french large white pigs at two distinct sampling times reveals strong signatures of convergent and divergent selection between the dam and sire lines. *Genetics Selection Evolution*, **55**, 13.

Bollback JP, York TL, Nielsen R (2008) Estimation of $2N_e s$ from temporal allele frequency data. *Genetics*.

Fages Antoine, Hanghøj Kristian, Khan Naveed, Gaunitz Charleen, Seguin-Orlando Andaine, Leonardi Michela, McCrory Constantz Christian, Gamba Cristina, Al-Rasheid Khaled A.S., Albizuri Silvia, Alfarhan Ahmed H., Allentoft Morten, Alquraishi Saleh, Anthony David, Baimukhanov Nurbol, Barrett James H., Bayarsaikhan Jamsranjav, Benecke Norbert, Bernáldez-Sánchez Eloísa, Berrocal-Rangel Luis, Biglari Fereidoun, Boessenkool Sanne, Boldgiv Bazartseren, Brem Gottfried, Brown Dorcas, Burger Joachim, Crubézy Eric, Daugnora Linas, Davoudi Hossein, Peter de BD, María de los Á de Cy de V-C, Deschler-Erb Sabine, Detry Cleia, Dill Nadine, Maria do MO, Dohr Anna, Ellingvåg Sturla, Erdenebaatar Diimaajav, Fathi Homa, Felkel Sabine, Fernández-Rodríguez Carlos, García-Viñas Esteban, Germonpré Mietje, Granado José D., Hallsson Jón H., Hemmer Helmut, Hofreiter Michael, Kasparov Aleksei, Khasanov Mutalib, Khazaeli Roya, Kosintsev Pavel, Kristiansen Kristian, Kubatbek Tabaldiev, Kuderna Lukas, Kuznetsov Pavel, Laleh Haeedeh, Leonard Jennifer A., Lhuillier Johanna, Liesau von Lettow-Vorbeck Corina, Logvin Andrey, Lõugas Lembi, Ludwig Arne, Luis Cristina, Arruda Ana Margarida, Marques-Bonet Tomas, Matoso Silva Raquel, Merz Victor, Mijiddorj Enkhbayar, Miller Bryan K., Monchalov Oleg, Mohaseb Fatemeh A., Morales Arturo, Nieto-Espinet Ariadna, Nistelberger Heidi, Onar Vedat, Pálsdóttir Albína H., Pitulko Vladimir, Pitskhelauri Konstantin, Pruvost Mélanie, Rajic Sikanjic Petra, Rapan Papeša Anita, Roslyakova Natalia, Sardari Alireza, Sauer Eberhard, Schafberg Renate, Scheu Amelie, Schibler Jörg, Schlumbaum Angela, Serrand Nathalie, Serres-Armero Aitor, Shapiro Beth, Sheikhi Seno Shiva, Shevnina Irina, Shidrang Sonia, Southon John, Star Bastiaan, Sykes Naomi, Taheri Kamal, Taylor William, Teegen Wolf-Rüdiger, Trbojević Vukičević Tajana, Trixl Simon, Tumen Dashzeveg, Undrakhbold Sainbileg, Usmanova Emma, Vahdati Ali, Valenzuela-Lamas Silvia, Viegas Catarina, Wallner Barbara, Weinstock Jaco, Zaibert Victor, Clavel Benoit, Lepetz Sébastien, Mashkour Marjan, Helgason Agnar, Stefánsson Kári, Barrey Eric, Willerslev Eske, Outram Alan K., Librado Pablo, Orlando Ludovic (2019) Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series. *Cell*, **177**, 1419–1435.e31.

Harris CR, Millman KJ, Walt SJ van der, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, Kerkwijk MH van, Brett M, Haldane A, Río JF del, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE (2020) Array programming with NumPy. *Nature*, **585**, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hui T-YJ (2014) NB: Maximum likelihood method in estimating effective population size from genetic data. *CRAN repository*.

Hui T-YJ, Burt A (2015) Estimating Effective Population Size from Temporally Spaced Samples with a Novel, Efficient Maximum-Likelihood Algorithm. *Genetics*, **200**, 285–293.

Long A, Liti G, Luptak A, Tenaillon O (2015) Elucidating the molecular architecture of adaptation via evolve and resequence experiments. *Nature Reviews Genetics*, **16**, 567–582.

Paris C, Servin B, Boitard S (2019b) compareHMM. *GitHub repository*.

Paris C, Servin B, Boitard S (2019a) Inference of selection from genetic time series using various parametric approximations to the wright-fisher model. *G3: Genes, Genomes,*

Genetics, **9**, 4073–4086.
Tataru P, Bataillon T, Hobolth A (2015) Inference under a wright-fisher model using an accurate beta approximation. *Genetics*, **201**, 1133–1141. https://doi.org/10.1534/genetics.115.179606
Uhl M, Navascués M de, Boitard S, Servin B (2024) Snt. *ForgeMIA repository*.