JOINT LIKELIHOOD-FREE INFERENCE OF THE NUMBER OF SELECTED SNPS AND THE SELECTION COEFFICIENT IN AN EVOLVING POPULATION

A PREPRINT

Yuehao Xu
Department of Applied Statistics
Johannes Kepler University Linz
Linz 4040, Austria
yuehao.xu@jku.at

Andreas Futschik
Department of Applied Statistics
Johannes Kepler University Linz
Linz 4040, Austria
andreas.futschik@jku.at

Ritabrata Dutta
Department of Statistics
University of Warwick
Coventry CV4 7AL, U.K.
ritabrata.dutta@warwick.ac.uk

September 20, 2022

ABSTRACT

With the exact likelihood often intractable, likelihood-free inference plays an important role in the field of population genetics. Indeed, several methodological developments in the context of Approximate Bayesian Computation (ABC) were inspired by population genetic applications. Here we explore a novel combination of recently proposed ABC tools that can deal with high dimensional summary statistics and apply it to infer selection strength and the number of selected loci for data from experimental evolution. While there are several methods to infer selection strength that operate on a single SNP level, our window based approach provides additional information about the selective architecture in terms of the number of selected positions. This is not trivial, since the spatial correlation introduced by genomic linkage leads to signals of selection also at neighboring SNPs. A further advantage of our approach is that we can easily provide an uncertainty quantification using the ABC posterior. Both on simulated and real data, we demonstrate a promising performance. This suggests that our ABC variant could also be interesting in other applications.

1 Introduction

Bayesian methods have become fashionable in the last couple of decades due to the considerable improvements in computational resources. Furthermore the development of simulation methods such as Markov Chain Monte Carlo (MCMC) permits to obtain approximate posteriors when exact computations would involve too complex integrals. Sometimes even an explicit evaluation of the likelihood is not possible. This is often true for population genetic models. Indeed, complex dependence patterns across genomic loci and the tree-shaped genealogical history of a sample leads to high dimensional nuisance parameters that often make exact likelihood computations cumbersome or even infeasible [4]. Thus, likelihood-free methods are popular in this field, since they only require to simulate from the underlying model [57]. In recent years, considerable effort has been put into developing methodology for likelihood-free inference (LFI) [34]. Approximate Bayesian Computation (ABC) ([4],[39]) and synthetic likelihood [55] are among the most popular approaches, and several variants have been proposed that differ in terms of simulation strategies, as well as the choice of distance measures and summary statistics. In any case, proper tuning is necessary in order to optimize the performance in a given application. For a more complete review on LFI see [3], [23] and [43].

Here, we develop an ABC based approach that is able to handle high-dimensional summary statistics in an application involving experimental evolution. Experiments in this context are conducted in order to better understand the genetic basis underlying evolutionary phenomena. For this purpose, populations of organisms are often exposed to a stressful condition in the laboratory. Over multiple generations, these organisms typically adapt. Genomic regions that are responsible for this adaptation will typically experience changes in allele frequencies that are larger than what would be expected by pure chance fluctuation. Using whole genome sequencing at multiple time points, data are generated to

search for such changes. The experiments are often replicated by starting with multiple genetically identical populations. Indeed, modern sequencing technologies can provide genome-wide data for samples of several populations at low cost. In Figure 1, the design of a typical experiment is explained. It should be pointed out that adaptation is also studied in natural populations. A famous example is the discovery of the genetic basis of the Tibetans' adaptation to a low oxygen atmosphere, see [42].

In recent years, different tools have been proposed to search genomic time series data for signals of selection. A composite-likelihood approach for evolve and resequence experiments (CLEAR) has been proposed by [25] to detect selected loci and to estimate their selection coefficients. A Bayesian approach called WFABC [18] provides estimates of the effective population size (N_e), selection and dominance coefficients under the Wright-Fisher model. Some statistics have also been proposed to test the presence of selection (e.g. [45], [13], [31], [38], [17], [30], [36], [15].) However, test statistics do not provide sposterior distributions for the selection coefficient. A recent comparison of methods for detecting and quantifying selection based on data from evolve and resequence experiments can be found in [51].

Several methods have also been developed to detect selection in natural populations. Selection can either be detected using data from present populations [7], or using also historic samples. Such past genomic information may consist of ancient DNA, or come from older samples conserved in biobanks. Historic samples contribute data from additional time points, but the underlying data generating process is different from experimental evolution. In such a context [35], [24] and [41] provide methods to infer selection along with allele ages from ancient genomic data based on Hidden Markov Models (HMM) using a diffusion approximation of the Wright-Fisher model. Further approaches in this direction are by [32], [14],[47] and [37]. For natural populations some authors also proposed methods that jointly infer demography and selection. The methods by [40], [26] and [19] use ABC for this purpose.

While the existing approaches typically operate on a SNP by SNP basis and assume at least locally a single target of selection, we operate on a window and also infer information about the local selective architecture. Indeed, our goal is to provide a methodology that permits to better understand how selection operates on a genomic window. For this purpose we use separate estimates of the selection coefficient for each SNP and for each pair of consecutive time points using a formula from [46]. These estimates are computed independently for each replicate and sorted according to average (across replicates) selection strength. This leads to large matrices that contain hundreds or even thousands of summary statistics, many of these may not be informative for our inferential goal. But we can handle these issues in ABC, since we consider a discrepancy measure which is able to deal with high dimensional summary statistics and automatically choose the most informative ones among them.

Following [21], we propose classification accuracy as the discrepancy measure to be used in our proposed ABC based inferential scheme. This is based on the simple intuition that the accuracy achieved by a classification rule is higher between two dissimilar data sets compared to two similar ones. With our high dimensional summary statistics, we focus on adaptive ℓ_1 penalized logistic regression as the classification rule, as it automatically chooses the most informative summary statistics while penalizing the non-informative ones. We then infer both the selection strength and the number of selected SNPs using simulated annealing ABC (SABC) [1] in conjunction with classification accuracy as the discrepancy measure, as it is computationally efficient and permits for parallel computing.

The rest of this article is organized as follows: In section 2, we explain our model, the data generating process, and the simulation setup. After briefly discussing the generic Approximate Bayesian Computation algorithm, section 3 introduces SABC. The rest of the section explains our choice of summary statistics and discrepancy measure. We present simulation results in section 4. A comparison with WFABC [18] and CLEAR [25] over a range of model parameter values is also provided. Section 5 contains a real data application. A discussion of our results and future directions in Section 6 concludes the manuscript.

2 Simulator model

Simulations are needed both to implement ABC and to evaluate the performance of our proposed approach. Figure 1 illustrates a typical evolve and resequence experiment that is used as template for our simulations. It starts with an initial population whose genetic composition is given by 189 founder haplotypes taken from a natural *D. simulans* population collected in Tallahassee, Florida, see [2] for further details. We consider experiments both for diploid and haploid organisms and use actual haplotypes from the experiment [2] mentioned above in our simulations. More specifically, our focus is on a window extending from position 363 to 49337 on chromosome 2L. This window contains a total of 500 SNPs. In the preprocessing step, we filter out SNPs with starting allele frequencies smaller than 0.1 and larger than 0.9 in order to reduce cases of early loss or fixation due to drift. The remaining 289 SNPs are used in the analysis. When simulating scenarios with positive selection, selected positions have been picked randomly among SNPs with starting allele frequencies between 0.1 and 0.5 since the detection power of any method will decrease when the starting allele frequency gets too close to zero or one.

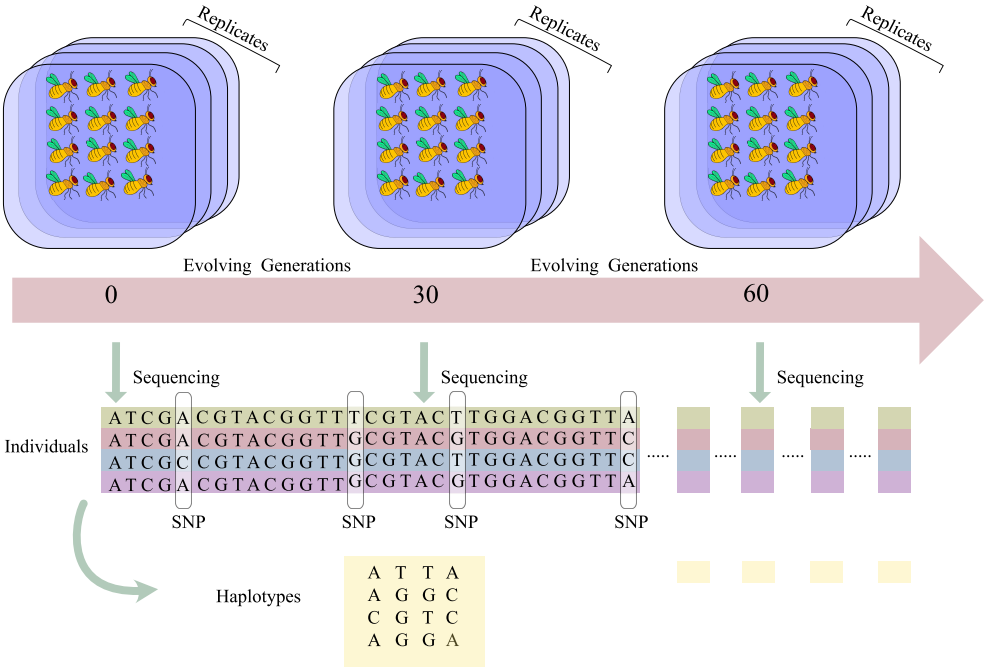

Figure 1: Design and data structure of a typical evolve and resequence experiment. Multiple replicate populations (blue) with controlled effective population size N_e evolve in the laboratory for 60 generations (pink). At the stated time points, a pool of DNA is extracted from each replicate population and sequenced. This leads to estimated allele frequencies at each SNP. Time is measured in terms of generation numbers.

The haplotype and resulting allele frequency dynamics is simulated via a Wright-Fisher model under different levels of selection. Details of the model are explained below. As default, we consider experiments with either ten or twenty identical replicated populations. Data for the replicates are generated by multiple simulation runs under identical parameter values and the same target(s) of selection. Obviously additional replicates provide more information, but do also lead to additional cost for the experimenter.

To get a comprehensive picture of the performance of our approach, we carried out simulations under different strengths of selection s, and numbers of selected targets. The values of further design parameters were also changed. Table 1 provides the default parameter values. Deviations from these values are indicated when discussing specific simulation results.

Since our ABC approach is computationally demanding, our results reflect the performance in 20 simulation runs (repetitions) per scenario. For haploid populations, we use our own simulation code, whereas for diploid simulations (with and without recombination), we use the software package *mimicrEE2* [52].

2.1 Wright-Fisher Model

Sewall Wright [56] and R.A. Fisher [16] proposed the so called Wright-Fisher model which is used extensively as a standard population genetic model. It is characterized by discrete and non-overlapping generations, constant population size with the new generation drawn by random sampling with replacement from the gene pool of the previous generation. In practice, deviations from these assumptions of the Wright-Fisher model are often encountered. Therefore the Wright-Fisher model is usually not applied with the census population size N but with the smaller effective population size N_e that produces the same amount of genetic drift as observed in the investigated population.

Overview of the parameters used for the simulations	
Parameters	Default Value
Chromosome	2L
Effective Population Size(N_e)	1000
Number of selected targets	1,2
Number of generations	60
Replicates	10, 20
Ploidy	1,2
Recombination	0 (negligible on chosen window)
Repetition	20
Number of SNPs	289
Founder Haplotypes	189

Table 1: Simulation parameter summary

Under neutrality, the sampling probabilities under a Wright-Fisher model are given by the haplotype frequencies in the parental generation. Thus expected frequencies stay constant across generations [11]. The haplotypes inherit the fitness of the SNPs they carry. The probability of drawing a haplotype is then proportional to its overall fitness. Assume that there are k SNPs with selection coefficients s_i ($1 \leq i \leq k$). Given k haplotypes $(h_1, h_2, \ldots, h_k)$, the fitness of haplotype h_j will be $w_j = 1 + \sum_i s_i(h_j)$. If haplotype h_j carries a neutral allele at position i, we have that $s_i(h_j) = 0$, and for beneficial and deleterious alleles we have $s_i(h_j) > 0$ and $s_i(h_j) < 0$ respectively. Assuming N_{Hap} haplotypes and a haploid organism, the haplotype frequencies $\mathbf{f_{t+1}}$ in generation $t+1$ are obtained via multinomial sampling using the frequencies $\mathbf{f_t}$ of the parental generation and the fitness vector $\mathbf{w} = (w_1, \ldots, w_p)$. More specifically,

$$\mathbf{f_{t+1}} \sim Multinomial[N_e, \frac{(f_{t,i}w_i)_{i=1}^k}{\sum_{j=1}^k f_{t,j}w_j}]/N_e. \tag{1}$$

Neutrality corresponds to the special case where all w_i are equal.

In a diploid population, each individual carries two haplotypes. Under the additive model, the contribution of SNP i with two types of alleles A and a to the fitness of individual J is the sum of the contributions over both alleles carried by the individual: $s_i(J) = \sum_{q=1}^2 s_{iq}(J)$. The overall contribution of a genomic region to the fitness of an individual is then obtained by summing over all SNPs within this region. In particular, for individual J we obtain $w(J) = 1 + \sum_i s_i(J)$. To form the next generation, N_e individuals are randomly drawn with replacement and drawing probabilities proportional to their fitness. The chosen individuals contribute one their segments, each with probability $1/2$.

In our simulations, we consider both haploid and diploid versions of the Wright-Fisher model.

3 Approximate Bayesian computation

A simple version of approximate Bayesian computation (ABC) known as rejection ABC was first applied to a genetics problem by [39]. This algorithm avoids the calculation of the likelihood, instead it simulates artificial data from the considered model using parameter drawn from the prior distribution. For each simulated parameter value, a distance is computed between observed and simulated summary statistics. Finally a randomly chosen parameter value is accepted as a sample point from the posterior, if the distance is smaller than a predetermined threshold > 0. This provides us with an approximation to the posterior indexed by the threshold and can be shown to converge to the true posterior when $\rightarrow 0$.

After the introduction of rejection ABC, several more efficient variants of ABC have been introduced following adaptive importance sampling strategies [33, 8]. They improve the approximation of the posterior distribution by reducing the

value of the threshold ϵ to 0 in an iterative manner. Here we use a variant that relies on simulated annealing, called simulated annealing approximate Bayesian computation (SABC) proposed in [1]. Our choice of simulate annealing ABC (SABC) is based on its computational efficiency due to the use of an optimal annealing schedule to decrease ϵ automatically, and its better suitability for high performance computing infrastructure as illustrated in [10].

3.1 Simulated Annealing ABC (SABC)

SABC proposed in [1] is an iterative algorithm, which starts by drawing an ensemble of particles that consist of parameter values θ drawn from the prior and the corresponding simulated model outputs x. In each step of iteration, new particles are simulated in the neighborhood of the previous ones using a time-dependent symmetric normal kernel. Each of the new particles replaces the corresponding previous one with probability $\min(1, \exp[-(u^* - u)/]$. Here u^* is a suitable distance between simulated output x^* and observed output y of the new particle, and u the same quantity computed from the previous simulated output x. The threshold for standard ABC algorithms plays the role of the temperature in the annealing algorithms that decreases over subsequent steps, so that in later steps only new particles are accepted that provide a better or similar quality approximation to y than the respective previous particle. Consideration of ϵ as temperature lets us to choose it in a decreasing manner such that the entropy of the system is optimized [44]. In statistical term this leads to a reduction of the information loss caused by rejections along the sampling process.

We notice that at each step of SABC, we obtain an approximation to the posterior distribution which improves as the number of steps increases and ϵ decreases. To ensure proper convergence to the posterior, the total number of iterative steps needs to be carefully chosen. The convergence of the posterior can be assessed through the Wasserstein distances [50] computed between the approximate posterior distributions at subsequent steps. When the values of the computed distances do not show any significant changes, we assume that the algorithm has converged. Based on pilot studies on simulated data sets with different selection coefficients (ranging from 0.02 to 0.09), we conclude that posterior convergence can be achieved in a robust manner after 25 step, a value we use for all studies in this work.

Some of the other tuning parameters (e.g. $\beta = 2, \delta = 0.2, \nu = 0.9$, acceptance rate cut-off = 0.001) of SABC have been chosen based on the recommendation in [1] and used as implemented in the python package ABCpy [9] which we used for inference. As a brief explanation, β governs the mixing in parameter space, δ governs the re-sampling steps, which represent the trade-off between bias reduction and effective sample size and ν governs the annealing speed. Furthermore, the initial ϵ should be chosen large enough to ensure that the approximate posterior at the initial step is similar to the prior distribution. As the classification accuracy explained in next Section [21] has been used as the distance measure in this work, we set ϵ as 1 which is the maximum value possible for classification accuracy. Further details of the algorithm can be found in [1].

Next we explain how informative summary statistics and the discrepancy measure between them was chosen for this specific problem.

3.2 Choice of summary statistics

The choice of informative summary statistics is important for any ABC algorithm. They serve to reduce the dimensionality of the original data while trying to capture as much information as possible. Previously proposed estimates and test statistics [29], [54], [49], [14], [45], [13], [31], [38], [17], [30], [36], [15] for the selection coefficient s seem to be the natural summary statistics in this case. Other quantities that are influenced by s, such as estimates of the effective population size (N_e) have also been used as summary statistics before, see [18]. The estimation of N_e both for natural and experimental populations has been investigated by several authors ([27], [53],[28]). Here, we tried several of these potential summary statistics and found that estimates of s obtained from logit-transformed allele frequencies being the most suitable ones. We explain these summary statistics and the intuitions behind them below.

When p_t is the allele frequency at generation $t \geq 0$, the allele frequency of a selected SNP in a haploid population can be modeled [46] as

$$\ln\left(\frac{p_t}{1-p_t}\right) = \ln\left(\frac{p_0}{1-p_0}\right) + st \tag{2}$$

in the absence of drift and sampling noise. The analogous equation for the diploid case is

$$\ln\left(\frac{p_t}{1-p_t}\right) = \ln\left(\frac{p_0}{1-p_0}\right) + \frac{s}{2}t. \tag{3}$$

In the presence of drift, this formula still provides a good approximation for intermediate allele frequencies when we have large enough populations and samples. This is because in such a setting the deterministic component dominates the noise. (Notice that in the experiments, we consider a setting similar to this by filtering out SNPs with low and high

starting frequencies.) Estimates of s are then computed for all SNPs and we sort the SNPs according to their average (across replicates) values. A further advantage of these estimates is that they are very fast to compute which helps to speed up ABC computations.

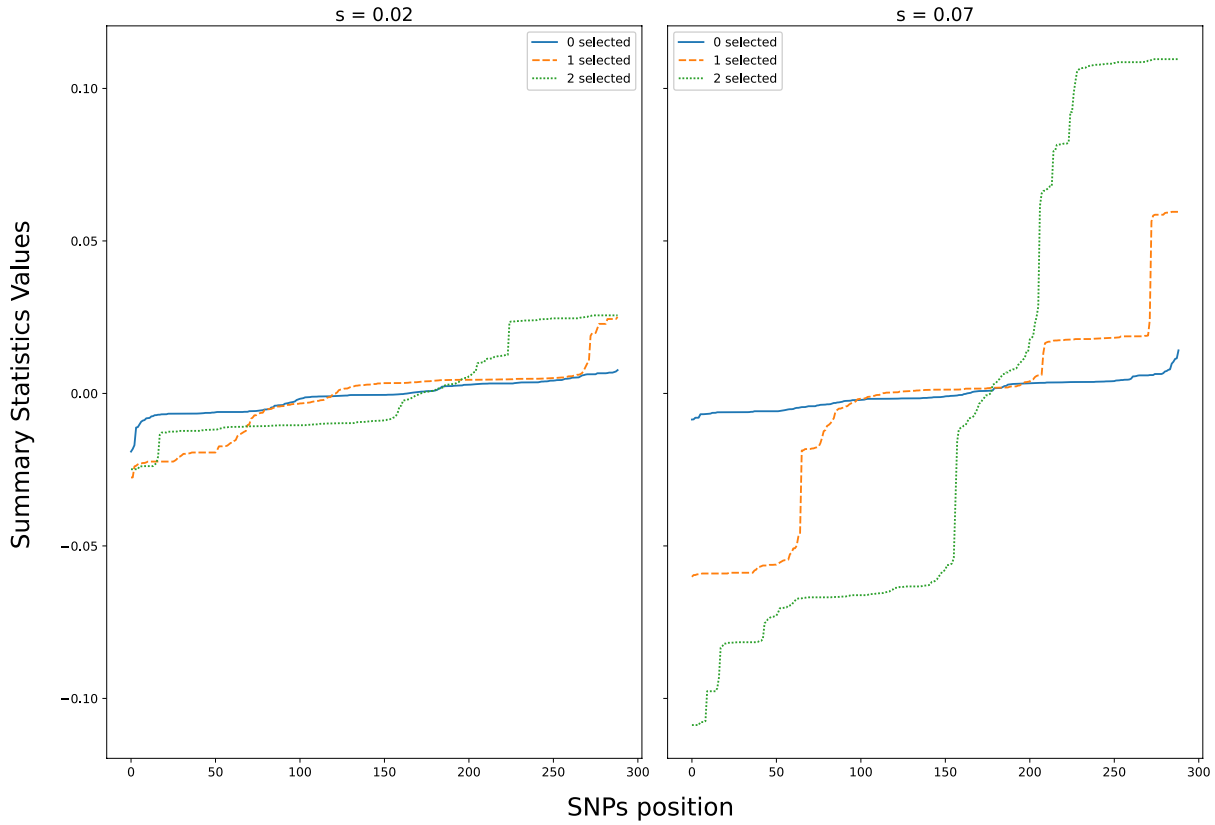

Figure 2: Average of the summary statistics over replicates for s = 0.02, 0.07. The Y-axis codes the values of summary statistics while the x-axis represents the ordered SNPs position. Different colors represent different numbers of selected loci. The data set involves $N_e = 1000$ diploid individuals that evolve for 60 generations. For all SNPs, the corresponding estimates of s are computed from the allele frequencies at generations 0 and 60. For the scenario with one selected SNP, the starting allele frequency is 0.191. For 2 selected targets the starting allele frequencies are 0.183 and 0.187 respectively.

Figures 2 and 3 provide examples that are based on one simulated data set for each considered scenario. They illustrate that our chosen summary statistics depend on both the selection strength, and also the number of loci (SNPs) under selection. Each data set has been generated using mimicrEE2. It involves ten replicate diploid populations with $N_e = 1000$, no recombination, and 60 generations. For 2 selected targets, the same selection strength (e.g. 0.02 or 0.07) is assumed for both. The heat maps in Figure 3 show that the estimates are more similar and lie around 0 when there is no selection ($ns = 0$). Under selection, larger estimates are present. These effects are more pronounced when s is large. The same patterns can also be seen in Figure 2, where the lines become steeper when the number of selected targets increases. The data sets used in the figures represent just one run of simulations, and are meant to illustrate typically encountered patterns. Notice, however, that especially for small values of s, there will be considerable random fluctuations between different runs and the resulting patterns in the summary statistics.

Notice that the vector of summary statistics considered here is high-dimensional. Depending on the selection strength and the number of selected targets, the number of informative SNPs varies. We therefore decided to choose the number of summary statistics used automatically as will be explained in the next subsection.

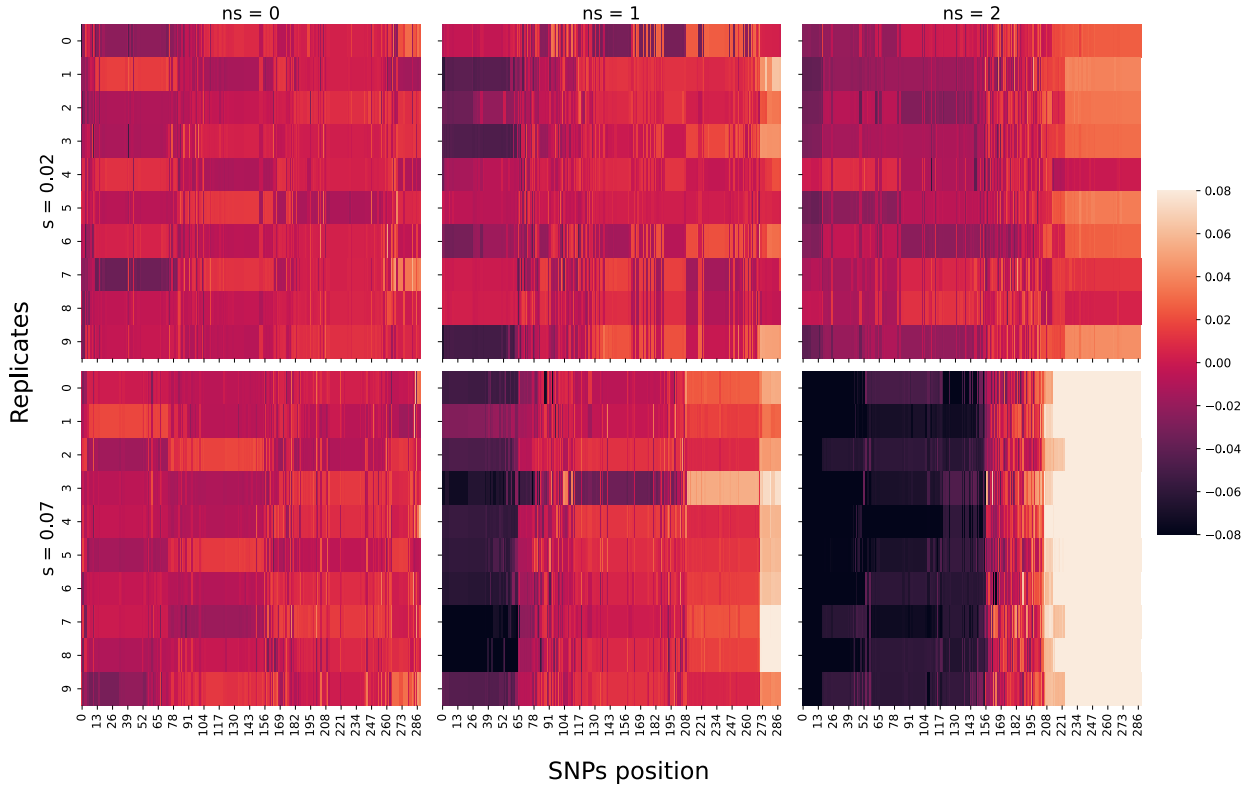

Figure 3: Heatmap of summary statistic values for different numbers of selected target. The x-axis represents the position of the SNPs and the y-axis represents replicates. The top maps are for $s = 0.02$, while the bottom maps are for $s = 0.07$. There are heatmaps for different numbers of selected targets (ns = 0 (left), 1 (center), 2(right column)).

3.3 Measuring discrepancy via adaptive ℓ_1 penalized logistics classification

With ABC the choice of the distance function plays an important role. The Euclidean distance between summary statistics computed from the observed and simulated data is commonly used for simple rejection ABC on low dimensional data. However, when it comes to high dimensional data, even their summary statistics are usually in high dimensions. Then the Euclidean distance will suffer from the curse of dimensionality and will be dominated by noise.

Several regularization techniques are available to deal with this curse of dimensionality. Methods such as LASSO [48], Smoothly Clipped Absolute Deviation (SCAD) [12], and Minimax Convex penalty (MCP) [58] use penalization to obtain sparse eistimates in high-dimensional classification and regression tasks. This idea was leveraged into recent applications of classification accuracy as discrepancy measure for ABC [21], where a penalized logistic regression was used as classification rule to estimate the classification accuracy. The penalty term in the logistic regression serves to automatically select the most informative summary statistics in a problem dependent manner. Developing on this, we propose to use the classification accuracy obtained via adaptive ℓ_1 penalized logistic regression for our problem in hand, which can automatically choose the most informative SNPs at each instance while computing the discrepancy. The details of this methodology are explained below.

To explain our approach, suppose that we compute summary statistics for each replicate population and label them with zero for the observed data $\{y^i_{obs},\ i = 1, \ldots, n\}$, and with one for the data simulated $\{y^i_\theta,\ i = 1, \ldots, n\}$ under parameter θ. This creates a labeled data set

$$D_\theta := \{(S(y^1_{obs}), 0), \ldots, (S(y^n_{obs}), 0), (S(y^1_\theta), 1) \ldots (S(y^n_\theta), 1)\},$$

where $S(\ )$ is the function computing summary statistics for each replicates. We can now learn a classification rule ϕ based on this data and consider the classification accuracy ξ of the classification rule ϕ,

$$\xi(\phi_\theta, D_\theta) = \frac{1}{2n}\sum_{i=1}^{n}([1-\phi_\theta(S(y)^i)] + \phi_\theta(S(y_\theta)^i)). \tag{4}$$

The larger the classification accuracy, the easier it is to distinguish between summary statistics computed from the observed and simulated data. Therefore the classification accuracy $\xi(\phi_\theta, D_\theta)$ can be interpreted as a discrepancy measure between observed and simulated summary statistics. The classification accuracy is estimated through leave one out cross-validation after training the classification rule on the data set, and this estimated classification accuracy is used as the discrepancy measure in ABC.

In our experimental setup, replicates present in the observed data corresponds to the number of data points (n in Equation 4). The sorted selection coefficients estimated for each SNP are considered as the features. Due to cost constraints, actual experiments for evolutionary genetics usually involve only around 10 or 20 replicates, whereas our considered window contains around 500 SNPs. In such a small sample with many features ($n >> m$) problem, we consider adaptive ℓ_1 penalized logistics regression [59] as the classification rule. In penalised logistic regression, the m regression coefficients corresponding to the m features present in the data are penalised by using ℓ_1 penalty. The penalty term of the adaptive ℓ_1 penalized logistics regression additionally involves adaptive weights that depend on the magnitude of an initial estimate β^* of the regression coefficients. A tuning parameter λ controls the penalty which can be chosen via cross-validation maximizing the classification accuracy, hence penalising the non-informative features in favor for the informative features. The resulting optimization problem becomes

$$\hat{\boldsymbol{\beta}}(\text{logistic}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\left(-y_i\left(\mathbf{x}_i^T\boldsymbol{\beta}\right) + \log\left(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}\right)\right) + \lambda\sum_{j=1}^{m}\hat{w}_j\,|\beta_j| \tag{5}$$

where $\hat{w}_j = \left(|\beta_j^*|\right)^{-\gamma}$ and γ is a positive constant. Due to the high dimensional nature of our problem, we use LASSO penalised logistic regression to estimate β^*. As some of the initial weights estimated by this procedure may be zero due to penalization, we add a small perturbation κ to the denominator while calculating the weights, leading to $\hat{w}_j = \left(|\beta_j^*| + \kappa\right)^{-\gamma}$. We chose κ and γ to be correspondingly 0.0001 and 1 in our computations following the existing convention in literature [22, 5]. The optimization problem in Equation 5 is numerically solved using glmnet [20].

4 Simulation Studies

We evaluate the performance of our approach by simulations under different scenarios. In particular we consider different selection strengths, population sizes, and numbers of replicates. We also simulate both haploid and diploid organisms, diploids also with recombination. Finally, we compare our method with two natural competitors. WFABC [18], as it relies also on an ABC implementation, and CLEAR [25], a particularly accurate method that uses composite likelihood to infer selection. In

In our SABC simulations, we used a uniform U[0, 0.2] prior for the selection coefficient, and a discrete uniform $DisreteUniform$[0, 3] prior for the number of selected targets. We collected samples of size 1,000 from the posterior.

With WFABC, we used the same prior for the selection coefficient. This method does not estimate the number of selected targets. Instead it produces posterior samples for every SNP. Since we do not know in practice which SNP is selected and which is just linked to a selected target, we take the maximum posterior mean over all SNPs as estimate of the selection coefficient. Since WFABC does not support replicates, we run it separately on all replicates and take the average of the resulting estimates. CLEAR, on the other hand, can combine information across multiple replicates but also provides estimates on a SNP by SNP basis. Again the maximum across all SNPs is taken as overall estimate. Further parameter choices with CLEAR were the maximum s value (0.2) and a step size of 0.001 for the grid search.

To quantify the performance when estimating the discrete number of selected targets, we introduce a quality score, which is calculated as follows: we assign a score of 1 if the estimate is correct, 0.5 if it differs by one, and 0 otherwise. We then average the scores over the simulation runs. We consider the posterior mode as our estimate of the selection coefficient.

Performance with haploid populations

In the haploid case, we consider 1000 individuals with $s \in [0.02, 0.05, 0.07, 0.09]$, and either 10 or 20 replicates. Figure 4 provides information on the performance of our estimates of s and the number of the selected targets. Here, the true

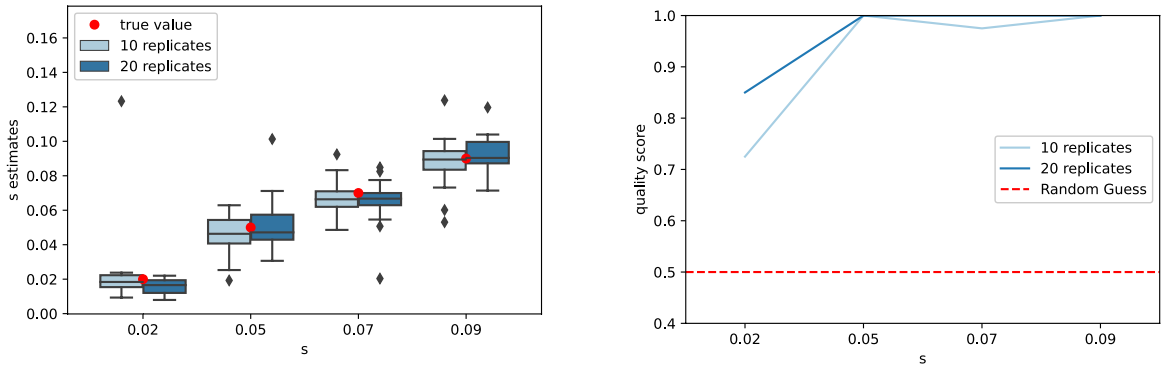

Figure 4: Performance in a scenario involving 1000 haploid individuals over a range of selection strengths for 10 and 20 replicate populations when we have one selected SNP. **Left:** The boxplots represent the posterior modes across our simulations. **Right:** Quality scores of the posterior mode of the number of selected targets.

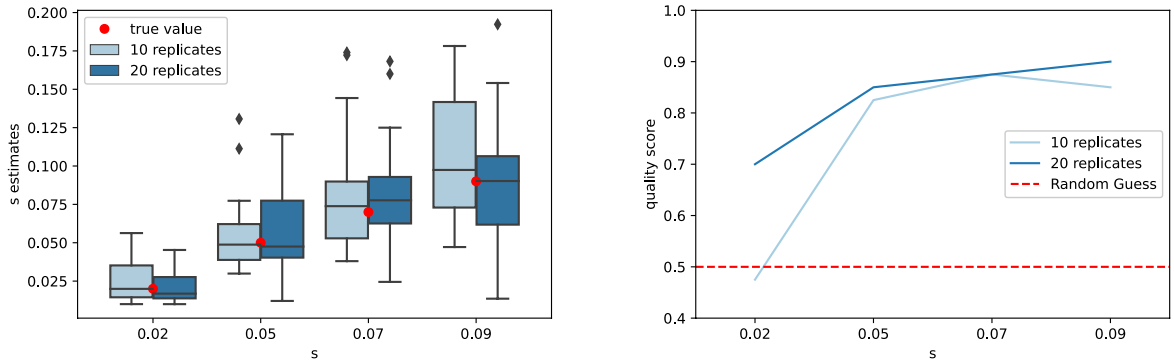

Figure 5: Performance in a scenario involving 1000 haploid individuals over a range of selection strengths for 10 and 20 replicate populations when we have two selected SNPs. The boxplots represent the posterior modes across our simulations. **Right:** Quality scores of the posterior mode of the number of selected targets.

number of selected SNPs is one. For each scenario, we conducted 20 simulation runs. The selection coefficient s is mostly estimated reliably, but there are occasional outliers. The quality scores for the number of the selected target are very close to the best possible score of one, especially when selection is not too weak. For $s = 0.02$, however, the performance of our estimates is worse than for larger values of s.

We next investigate a scenario with two selected SNPs. The other parameter values are chosen as before. We applied the same selection strength to both SNPs for two selected targets. As shown in figure 5, the variance of the estimates increases, and the quality score decreases compared to the case of one selected target. Nevertheless, we still achieve reasonable accuracy.

Comparison of our approach with WFABC and CLEAR for haploid individuals

Figure 6 provides a comparison of our approach with two competitors, CLEAR and WFABC. Since the competing methods cannot estimate the number of selected SNPs, we consider scenarios with one selected SNP and focus on the performance when estimating s. Our results indicate, that WFABC provides biased estimates, in particular when

the selection strength s is low. CLEAR estimates s with similar accuracy as our method. Since previous comparisons **(author?)** [51] suggest that CLEAR performs particularly well, we do not expect further alternative approaches to outperform our method. While our estimates are slower to compute than those obtained via CLEAR, they also provides information on the number of selected targets. Furthermore they come with a quantification of uncertainty in terms of the posterior.

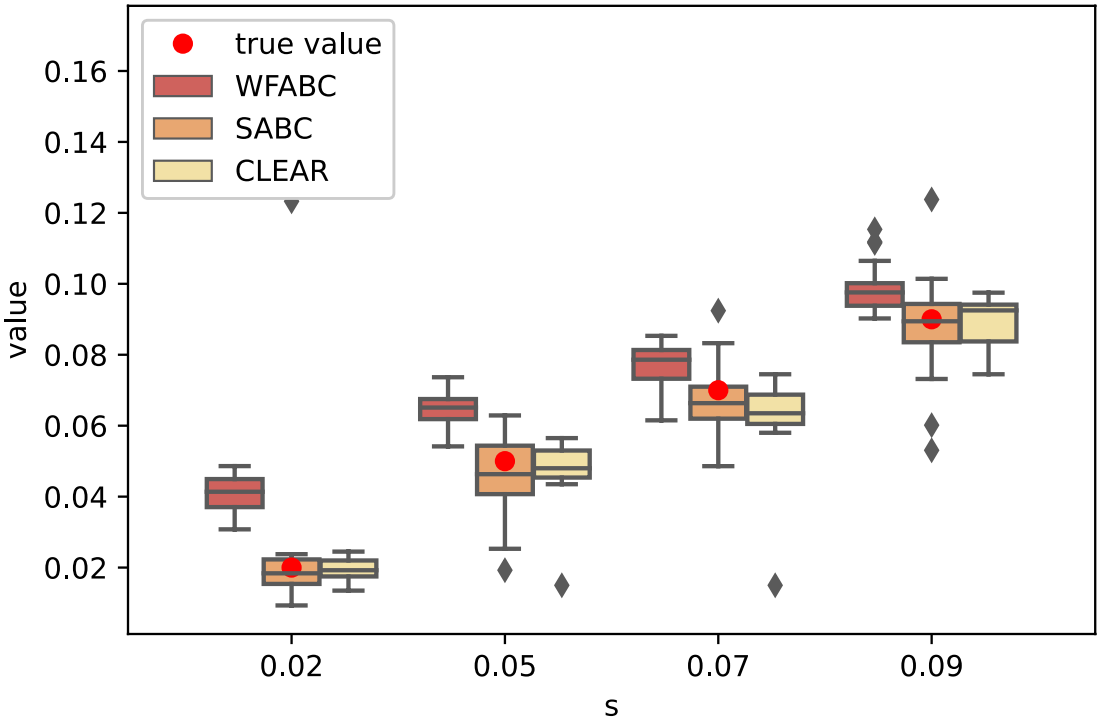

Figure 6: Comparison between our method (SABC), WFABC [18] and CLEAR [25] for haploid individuals. Twenty simulations of an experiment involving ten replicate populations have been carried out for each value of s. The further parameters are as in Table 1, with one selected target and allele frequency data available every 10 generations.

Influence of ploidy and recombination on the accuracy

We consider experiments where ten replicate populations of 1000 diploid individuals and 2000 haploid individuals evolved for 60 generations. The true s is 0.07 for all three scenarios. The selected SNP was chosen the same throughout all 40 experiments for both the diploid and haploid case. The recombination rate is 2.5 cM/Mb, and the dominance coefficient h is assumed to be 0.5 which leads to a so called additive model.

The results in figure 7 are for a scenario with one selected target, whereas figure 8 considers the case of tho selected SNPs. We observe that without recombination, the estimates have a smaller variance compared to the case where recombination is present. This suggests that recombination makes the estimation more difficult since it introduces additional noise. With respect to the number of selected targets, the achieved quality scores suggest that the haploid case is easier compared to diploid.

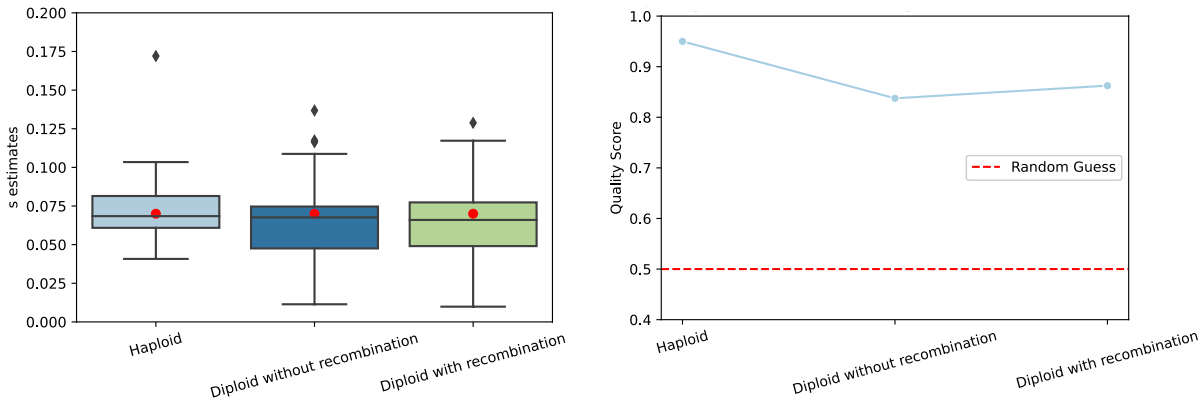

Figure 7: Performance of our estimates under three scenarios: Haploid, Diploid with recombination (recombination rate : 2.5 cM/Mb), Diploid without recombination; the true selection strength is 0.07 (red dot) and one selected target is present in the data. All three scenarios include 40 replicates. **Left:** Estimates of s **Right:** Quality scores of estimates of the number of selected targets.

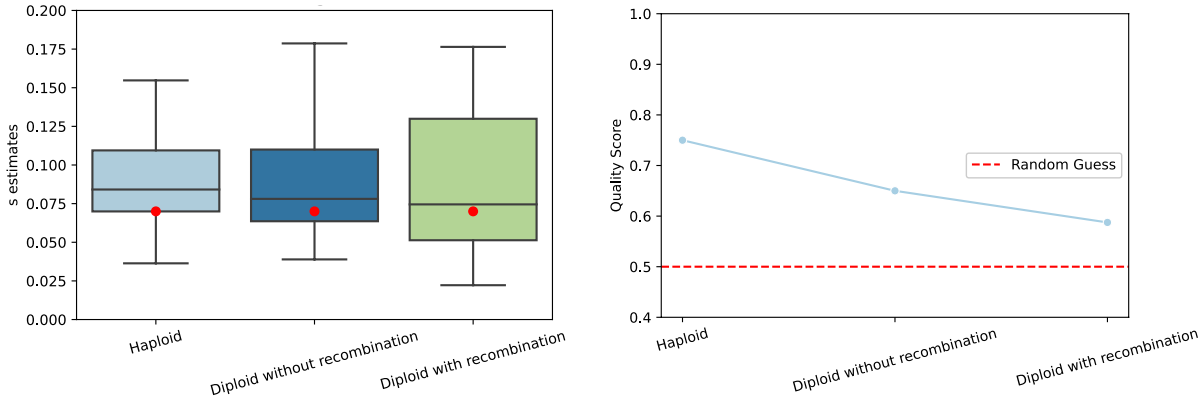

Figure 8: Performance of our estimates under three scenarios: Haploid, Diploid with recombination (recombination rate : 2.5 cM/Mb), Diploid without recombination; the true selection strength is 0.07 (red dot) and two selected targets are present in the data. All three scenarios include 40 replicates. **Left:** Estimates of s **Right:** Quality scores of estimates of the number of selected targets.

5 Inference on real dataset

Here we illustrate the application of our method on real data. More specifically, we take outcrossing yeast data from [6]. The data set contains allele frequencies at four time points (generations {0,180,360,540}) from 12 replicate populations. We remove the allele frequencies for the last time point because most SNPs get lost or fixed at generation 540.

On chromosome 11, we consider both a small window of 140 SNPs (positions from 3049 to 37862) and a larger window containing 500 SNPs for this window (positions from 3049 to 88212). We confirmed the presence of selection in this region using the existing methods of [25] and [46]. Selection seems to be weak however, as the largest estimate of s from CLEAR was 0.009. Additionally CLEAR provides also the estimate $\hat{N}_e \approx 1000$ for the effective population size. In order to apply our method to this window, we need simulated allele frequencies. For this purpose, we run our simulator model with the estimated value of N_e and the known starting haplotypes. As selection has already been confirmed for the considered region, we take a Discrete Uniform [1,2,3] prior for the number of selected targets. From the accepted parameter values of s, we obtain kernel density estimates of the posterior density.

Figures 9 and 10 show that the reference estimate obtained from CLEAR lies close to the posterior mode both for the smaller and the larger window, suggesting that both methods provide similar estimates. The posterior sample produced by our method, however, also describes the uncertainty associated with the estimates. Also, our method suggests the presence of one selected target for the smaller window, and two targets of selection for the larger one. See again Figures 9 and 10, for more detailed results.

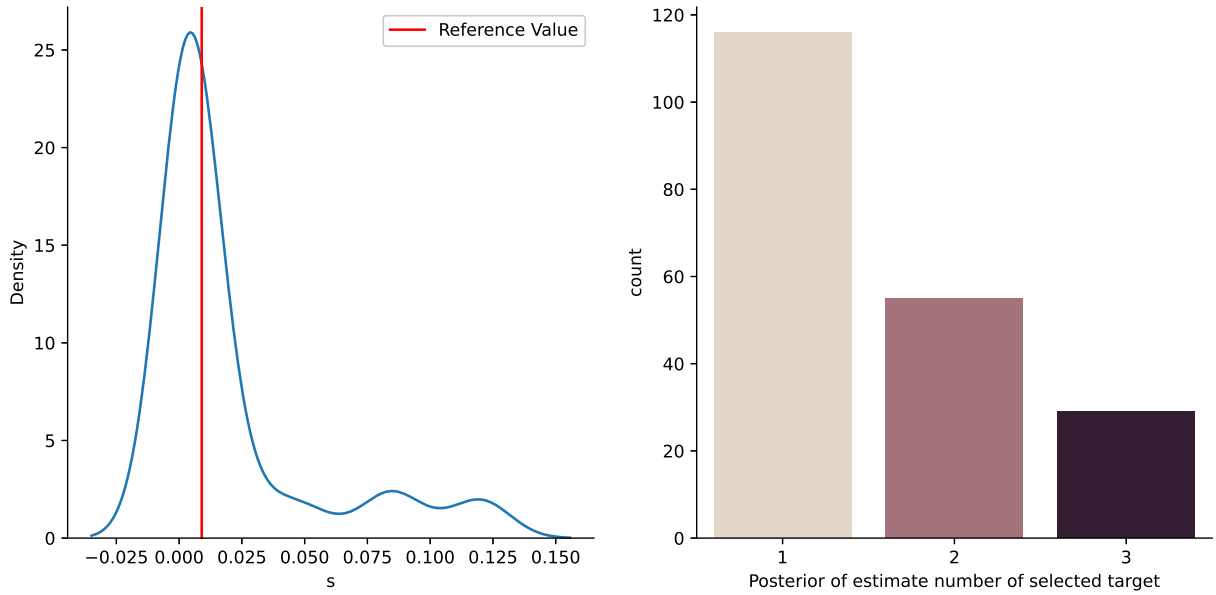

Figure 9: Data for a 140 SNP window from experiment on yeast by [6]. **Left**: Posterior from our estimate. The red line represents the selection coefficient estimated using [25] and [46]. **Right**: posterior for the number of selected targets.

6 Discussion

For experimental evolution data, methods have been proposed that test for selection. Estimates for the selection coefficient s are also available. These methods typically operate on a SNP by SNP basis and do not distinguish between the selected targets and hitchhiking phenomena at neighboring positions due to linkage. Here we propose a different window-based approach that estimates the number of selected targets and their selection coefficient under the assumption that both targets have the same strength of selection for a given window in the genome. Our approach relies on high dimensionsional summary statistics and a novel combination of ABC tools including SABC[1] and penalized logistic regression.

When estimating s, our method is more accurate than WFABC[18], a method that uses simple rejection ABC. We achieved a similar level as accuracy as CLEAR, a method that uses composite likelihood computations and is particularly

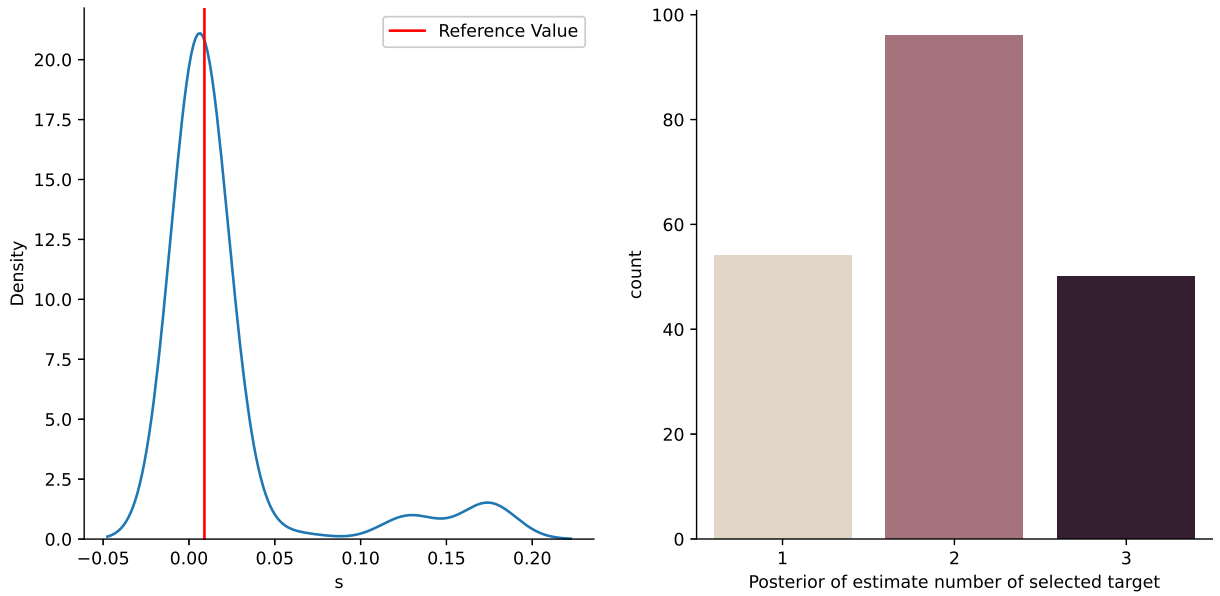

Figure 10: Data for a 500 SNP window from experiment on yeast by [6]. **Left**: Posterior obtained by our estimate. The red line represents the selection coefficient estimated using [25] and [46]. **Right**: posterior for the number of selected targets.

accurate. However, an advantage of our method is that we provide a posterior estimate to assess uncertainty. Another feature unique to our method is that we also estimate the number of selected SNPs in the window. We believe this feature will be of interest to the field of experimental evolution. A current limitation of our work is that our estimates of the number of selected targets are not very accurate, if there are more than two targets. Indeed, with a larger number of selected SNPs, the algorithm often underestimates the number of selected positions. We plan to explore this issue further in future work. Further plans are to also consider situations with dominance effects, as well as additional modes of selection, such as truncating and stabilizing selection.

Acknowledgements

RD is funded by pdfRC (grant nos. EP/V025899/1, EP/T017112/1) and NERC (grant no. NE/T00973X/1). We acknowledge helpful comments by Yuexuan Wang and Marta Pelizzola.

Supporting Information

The source code can be found at `https://github.com/YuehaoXu96/Joint-Inference`. The D.simulan data is available at [2]. Outcrossing yeast data available at [6].

References

[1] Carlo Albert, Hans R Künsch, and Andreas Scheidegger. A simulated annealing approach to approximate bayes computations. *Statistics and computing*, 25(6):1217–1232, 2015.

[2] Neda Barghi, Raymond Tobler, Viola Nolte, Ana Marija Jakšić, François Mallard, Kathrin Anna Otte, Marlies Dolezal, Thomas Taus, Robert Kofler, and Christian Schlötterer. Genetic redundancy fuels polygenic adaptation in drosophila. *PLoS biology*, 17(2):e3000128, 2019.

[3] Mark A Beaumont. Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, 41:379–406, 2010.

[4] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[5] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[6] Molly K Burke, Gianni Liti, and Anthony D Long. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of saccharomyces cerevisiae. *Molecular biology and evolution*, 31(12):3228–3239, 2014.

[7] Michael DeGiorgio, Christian D Huber, Melissa J Hubisz, Ines Hellmann, and Rasmus Nielsen. Sweepfinder2: increased sensitivity, robustness and flexibility. *Bioinformatics*, 32(12):1895–1897, 2016.

[8] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and computing*, 22(5):1009–1020, 2012.

[9] Ritabrata Dutta, Marcel Schoengens, Jukka-Pekka Onnela, and Antonietta Mira. Abcpy: A user-friendly, extensible, and parallel library for approximate bayesian computation. In *Proceedings of the platform for advanced scientific computing conference*, pages 1–9, 2017.

[10] Ritabrata Dutta, Marcel Schoengens, Lorenzo Pacchiardi, Avinash Ummadisingu, Nicole Widmer, Pierre Künzli, Jukka-Pekka Onnela, and Antonietta Mira. Abcpy: A high-performance computing perspective to approximate bayesian computation. *Journal of Statistical Software*, 100(7):1–38, 2021.

[11] Warren John Ewens. *Mathematical population genetics: theoretical introduction*, volume 1. Springer, 2004.

[12] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[13] Justin C Fay and Chung-I Wu. Hitchhiking under positive darwinian selection. *Genetics*, 155(3):1405–1413, 2000.

[14] Alison F Feder, Sergey Kryazhimskiy, and Joshua B Plotkin. Identifying signatures of selection in genetic time series. *Genetics*, 196(2):509–522, 2014.

[15] Anna Ferrer-Admetlla, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular biology and evolution*, 31(5):1275–1291, 2014.

[16] Ronald Aylmer Fisher. *The genetical theory of natural selection*. Oxford University Press, 1930.

[17] Matthieu Foll and Oscar Gaggiotti. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics*, 180(2):977–993, 2008.

[18] Matthieu Foll, Hyunjin Shim, and Jeffrey D Jensen. Wfabc: a w right–f isher abc-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular ecology resources*, 15(1):87–98, 2015.

[19] Christelle Fraïsse, Iva Popovic, Clément Mazoyer, Bruno Spataro, Stéphane Delmotte, Jonathan Romiguier, Etienne Loire, Alexis Simon, Nicolas Galtier, Laurent Duret, et al. Dils: Demographic inferences with linked selection by using abc. *Molecular Ecology Resources*, 21(8):2629–2644, 2021.

[20] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[21] Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, 2018.

[22] Sung Won Han, Gong Chen, Myun-Seok Cheon, and Hua Zhong. Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. *Journal of the American Statistical Association*, 111(515):1004–1019, 2016.

[23] Florian Hartig, Justin M Calabrese, Björn Reineking, Thorsten Wiegand, and Andreas Huth. Statistical inference for stochastic simulation models–theory and application. *Ecology letters*, 14(8):816–827, 2011.

[24] Zhangyi He, Xiaoyang Dai, Mark Beaumont, and Feng Yu. Maximum likelihood estimation of natural selection and allele age from time series data of allele frequencies. *bioRxiv*, page 837310, 2020.

[25] Arya Iranmehr, Ali Akbari, Christian Schlötterer, and Vineet Bafna. Clear: Composition of likelihoods for evolve and resequence experiments. *Genetics*, 206(2):1011–1023, 2017.

[26] Parul Johri, Brian Charlesworth, and Jeffrey D Jensen. Toward an evolutionarily appropriate null model: jointly inferring demography and purifying selection. *Genetics*, 215(1):173–192, 2020.

[27] Ágnes Jónás, Thomas Taus, Carolin Kosiol, Christian Schlötterer, and Andreas Futschik. Estimating the effective population size from temporal allele frequency changes in experimental evolution. *Genetics*, 204(2):723–735, 2016.

[28] Per Erik Jorde and Nils Ryman. Unbiased estimator for genetic drift and effective population size. *Genetics*, 177(2):927–935, 2007.

[29] John K Kelly and Kimberly A Hughes. Pervasive linked selection and intermediate-frequency alleles are implicated in an evolve-and-resequencing experiment of drosophila simulans. *Genetics*, 211(3):943–961, 2019.

[30] Yuseob Kim and Rasmus Nielsen. Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167(3):1513–1524, 2004.

[31] Yuseob Kim and Wolfgang Stephan. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777, 2002.

[32] Miguel Lacerda and Cathal Seoighe. Population genetics inference for longitudinally-sampled mutants under strong selection. *Genetics*, 198(3):1237–1250, 2014.

[33] Maxime Lenormand, Franck Jabot, and Guillaume Deffuant. Adaptive approximate bayesian computation for complex models. *Computational Statistics*, 28(6):2777–2796, 2013.

[34] Jarno Lintusaari, Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Fundamentals and recent developments in approximate bayesian computation. *Systematic biology*, 66(1):e66–e82, 2017.

[35] Anna-Sapfo Malaspinas, Orestis Malaspinas, Steven N Evans, and Montgomery Slatkin. Estimating allele age and selection coefficient from time-serial data. *Genetics*, 192(2):599–607, 2012.

[36] Rasmus Nielsen, Scott Williamson, Yuseob Kim, Melissa J Hubisz, Andrew G Clark, and Carlos Bustamante. Genomic scans for selective sweeps using snp data. *Genome research*, 15(11):1566–1575, 2005.

[37] Cyriel Paris, Bertrand Servin, and Simon Boitard. Inference of selection from genetic time series using various parametric approximations to the wright-fisher model. *G3: Genes, Genomes, Genetics*, 9(12):4073–4086, 2019.

[38] Pavlos Pavlidis, Daniel Živković, Alexandros Stamatakis, and Nikolaos Alachiotis. Sweed: likelihood-based detection of selective sweeps in thousands of genomes. *Molecular biology and evolution*, 30(9):2224–2234, 2013.

[39] Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[40] Andrew M Sackman, Rebecca B Harris, and Jeffrey D Jensen. Inferring demography and selection in organisms characterized by skewed offspring distributions. *Genetics*, 211(3):1019–1028, 2019.

[41] Joshua G Schraiber, Steven N Evans, and Montgomery Slatkin. Bayesian inference of natural selection from allele frequency time series. *Genetics*, 203(1):493–511, 2016.

[42] Tatum S Simonson, Yingzhong Yang, Chad D Huff, Haixia Yun, Ga Qin, David J Witherspoon, Zhenzhong Bai, Felipe R Lorenzo, Jinchuan Xing, Lynn B Jorde, et al. Genetic evidence for high-altitude adaptation in tibet. *Science*, 329(5987):72–75, 2010.

[43] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

[44] Wolfgang Spirkl and Harald Ries. Optimal finite-time endoreversible processes. *Physical Review E*, 52(4):3485, 1995.

[45] Fumio Tajima. Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105(2):437–460, 1983.

[46] Thomas Taus, Andreas Futschik, and Christian Schlötterer. Quantifying selection with pool-seq time series data. *Molecular biology and evolution*, 34(11):3023–3034, 2017.

[47] Jonathan Terhorst, Christian Schlötterer, and Yun S Song. Multi-locus analysis of genomic time series data from experimental evolution. *PLoS genetics*, 11(4):e1005069, 2015.

[48] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[49] Hande Topa, Ágnes Jónás, Robert Kofler, Carolin Kosiol, and Antti Honkela. Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics*, 31(11):1762–1770, 2015.

[50] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[51] Christos Vlachos, Claire Burny, Marta Pelizzola, Rui Borges, Andreas Futschik, Robert Kofler, and Christian Schlötterer. Benchmarking software tools for detecting and quantifying selection in evolve and resequencing studies. *Genome biology*, 20(1):1–11, 2019.

[52] Christos Vlachos and Robert Kofler. Mimicree2: Genome-wide forward simulations of evolve and resequencing studies. *PLoS computational biology*, 14(8):e1006413, 2018.

[53] Robin S Waples. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, 121(2):379–391, 1989.

[54] R Axel W Wiberg, Oscar E Gaggiotti, Michael B Morrissey, and Michael G Ritchie. Identifying consistent allele frequency differences in studies of stratified populations. *Methods in ecology and evolution*, 8(12):1899–1909, 2017.

[55] Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104, 2010.

[56] Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

[57] Xiguo Yuan, David J Miller, Junying Zhang, David Herrington, and Yue Wang. An overview of population genetic data simulation. *Journal of Computational Biology*, 19(1):42–54, 2012.

[58] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[59] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.