Article – Methods

Bait-ER: a Bayesian method to detect targets of selection in Evolve-and-Resequence experiments

Carolina Barata[1,†], Rui Borges[2,†] and Carolin Kosiol[1,2,*]

1. Centre for Biological Diversity, University of St Andrews, St Andrews, Fife KY16 9TH, UK 2. Institute of Population Genetics, Vetmeduni Vienna, Veterinärplatz 1, 1210 Wien, Austria

† These authors contributed equally and will be putting their name first on the citation in their CVs.

* Corresponding author: ck202@st-andrews.ac.uk

Bait-ER: a Bayesian method to detect targets of selection in Evolve-and-Resequence experiments

Abstract

For over a decade, experimental evolution has been combined with high-throughput sequencing techniques in so-called Evolve-and-Resequence (E&R) experiments. This allows testing for selection in populations kept in the laboratory under given experimental conditions. However, identifying signatures of adaptation in E&R datasets is far from trivial, and it is still necessary to develop more efficient and statistically sound methods for detecting selection in genome-wide data. Here, we present Bait-ER – a fully Bayesian approach based on the Moran model of allele evolution to estimate selection coefficients from E&R experiments. The model has overlapping generations, a feature that describes several experimental designs found in the literature. We tested our method under several different demographic and experimental conditions to assess its accuracy and precision, and it performs well in most scenarios. However, some care must be taken when analysing specific allele trajectories, particularly those where drift largely dominates and starting frequencies are low. We compare our method with other available software and report that ours has generally high accuracy even for very difficult trajectories. Furthermore, our approach avoids the computational burden of simulating an empirical null distribution, outperforming available software in terms of computational time and facilitating its use on genome-wide data.

We implemented and released our method in a new open-source software package that can be accessed at `https://github.com/mrborges23/Bait-ER`.

Key-words: targets of selection, E&R, pool-seq, selection coefficients, Moran model, Bayesian inference

1 Introduction

Natural selection is a very complex process that can dramatically alter phenotypes and genotypes over remarkably short timescales. Researchers have successfully tested theoretical predictions and collected evidence for how strong laboratory selection acting on phenotypes can be. However, it is not as straightforward to measure selection acting on the genome. There are many confounding factors that can lead to spurious results. This is particularly relevant if we are interested in studying how experimental populations adapt to laboratory conditions within tens of generations, in which case we need to take both experiment- and population-related parameters into account.

Regardless of said difficulties, numerous experimental evolution studies, where populations are exposed to a controlled laboratory environment for some number of generations (Kawecki et al., 2012), have made remarkable discoveries on the genomic architecture of adaptation. Examples of long-term experimental evolution studies include those on yeast (Burke et al., 2014), red flour beetles (Godwin et al., 2017) and fruit flies (Turner et al., 2011; Debelle et al., 2017). By combining experimental evolution with next-generation sequencing, Evolve-and-Resequence studies (E&R, **fig. 1**) can shed light on the genetic basis of short-term adaptation. The E&R set-up allows for describing the divergence between experimental treatments while accounting for variation among replicate populations (Schlötterer et al., 2015). In other words, to find signatures of selection, one must not only monitor allele frequency changes throughout the experiment but also search for consistency across replicates. Moreover, it is often the case that experimental populations are sampled and pooled for genome sequencing. Sequencing pooled samples of individuals (pool-seq) is cost-effective and produces largely accurate estimates of population allele frequencies (Futschik and Schlötterer, 2010). Thus, statistical methods developed for E&R studies are especially useful if our aim is to find signatures of selection across the genome. Notably so when we investigate allele frequency trajectories originating from pooled samples.

In contrast to E&R experiments in bacteria, those in sexual eukaryotes aim at describing adaptation due to standing genetic variation rather than that caused by new advantageous mutations. E&R time series datasets are particularly well-suited to fully characterise allele frequency trajectories. Several statistical approaches have been proposed to analyse these data and detect signatures of selection across the genome. A few such methods consider allele frequency changes between two time points. These simply identify those loci where there is a consistent difference in frequency between time points. One such approach is the widely-used Cochran-Mantel-Haenszel (CMH) test (Cochran, 1954). Such tests are often preferred since they are very fast, which makes them suitable for genome-wide datasets. Other approaches employ methods that allow for more than two time points: for example, Wiberg et al. (2017) used generalised linear models, and introduced a quasi-binomial distribution for the residual

error; and Topa et al. (2015) employed Gaussian Process models in a Bayesian framework to test for selection while accounting for sampling and sequencing noise. While the latter methods use more sophisticated statistical approaches, they remain descriptive and empirical with respect to underlying evolutionary processes. In contrast, mechanistic approaches explicitly model evolutionary forces, such as genetic drift and selection. Such models have the advantage that they can properly account for drift, which may generate allele frequency changes that can easily be mistaken for selection. Indeed, this is usually the case for E&R experimental populations with low effective population sizes (N_e), where genetic drift is the main evolutionary force determining the fate of most alleles.

To our knowledge, three main mechanistic methods have thus far been developed: Wright-Fisher Approximate Bayesian Computation (WFABC, Foll et al. (2015)), Composition of Likelihoods for E&R experiments (CLEAR, Iranmehr et al. (2017)) and LLS (Linear Least Squares, Taus et al. (2017)). These methods differ in how they model drift and selection, the inferential approach to estimate selection coefficients, the hypothesis testing strategy, and the extent to which they consider specific experimental conditions (**table 1**). WFABC employs an ABC approach which uses summary statistics to compare simulated and real data. This method jointly infers the posterior of both N_e and the selection coefficient at some locus in the genome using allele frequency trajectory simulations. It performs simulations until both real and simulated summary statistics agree to a certain predefined scale. This makes WFABC computationally intensive. CLEAR computes maximum-likelihood estimates of selection parameters using a hidden Markov model tailored for small population sizes. LLS assumes that allele frequencies vary linearly with selection coefficients such that the slope provides the coefficient estimate. Although all three methods have been shown to accurately estimate selection coefficients, they rely heavily on empirical parameter distributions to perform hypothesis testing: (i) WFABC is highly dependent on the priors used to simulate those trajectories; (ii) CLEAR relies on genome-wide simulations to calculate an empirical likelihood-ratio statistic to assess significance; and (iii) LLS computes an empirical distribution of p-values simulated under neutrality. Additionally, the three software vary substantially on computational effort. Therefore, currently available methods are still limited in their use for genome-wide hypothesis testing.

Here, we propose a new Bayesian inference tool – Bait-ER – to estimate selection coefficients in E&R time series data. It is suitable for large genome-wide polymorphism datasets and particularly useful for small experimental populations. We show that our method is faster than other available software (when accounting for hypothesis testing) while still performing accurately in some particularly difficult scenarios.

2 New Approaches

E&R experiments produce a remarkable amount of data, namely allele counts for thousands to millions of loci. We created a Bayesian framework to infer and test for selection at an individual locus that is based on the Moran model. The Moran model is especially useful for studies that have overlapping generations, such as insect cage experimental designs (**fig. 1**). Such cage experiments are easier to maintain in the lab and allow for larger experimental population sizes avoiding potential inbreeding depression and crashing populations. Furthermore, Bait-ER combines modelling the evolution of an allele that can be under selection while accounting for sampling noise to do with pooled sequencing and finite sequencing depth. Our method takes allele count data in the widely-used sync format (Kofler et al., 2011) as input. Each locus is described by allele counts per time point and replicate population. The algorithm implemented includes the following key steps:

1. Bait-ER calculates the virtual allele frequency trajectories accounting for N_e that is provided by the user. This step includes a binomial sampling process that corrects for pool-seq-associated sampling noise.

2. The log posterior density of σ is calculated for a given grid of σ-values. This step requires repeatedly assessing the likelihood function (equation 3 in section 4).

3. The log posterior values obtained in the previous step are fitted to a gamma surface (details on surface fitting can be found in **supplementary fig. S1**).

4. Bait-ER returns a set of statistics that describe the posterior distribution of σ per locus. In particular, the average σ and the log Bayes Factor (BF) are the most important quantities. In this case, BFs test the hypothesis that σ is different from 0. Bait-ER also returns the posterior shape and rate parameter values, α and β, respectively. These can be used to compute other relevant statistics (e.g., credible intervals, variance).

As our new approach was implemented in a Bayesian framework, it allows for measuring uncertainty associated with inference, for it gives posterior distributions of any selection parameters. Bait-ER jointly tests for selection and estimates selection parameters contrary to other state-of-the-art methods. It does not rely on empirical or simulation-based approaches that might be computationally intensive, and it properly accounts for specific shortcomings of E&R experimental design. Bait-ER performs well even for trajectories simulated under complex demographic scenarios.

3 Results and Discussion

3.1 Prior fitting with Bait-ER

Bait-ER employs a Bayesian approach outlined in the New Approaches section and described in further detail in the Methods section. Bayesian model fitting depends on the prior distribution implemented and requires further testing. Bait-ER uses a gamma prior for which the shape α and rate β parameters have to be defined beforehand. We tested the impact of uninformative ($\alpha = \beta = 0.001$) and informative ($\alpha = \beta = 10^5$) gamma priors on the posterior distribution of σ under standard (60x coverage, 5 time points and 5 replicates) and sparse (20x coverage, 2 time points and 2 replicates) E&R experiments. Our results show that the prior parameters have virtually no impact on the posterior estimates when $\alpha = \beta < 100$ (**fig. 2** and **supplementary fig. S1**), and thus, by default, Bait-ER sets both prior parameters to 0.001.

Calculating the posterior distribution of σ is a computationally intensive step because it requires solving the exponential Moran matrix for several σ-values. To reduce the number of times Bait-ER assesses the log-posterior, we fit the posterior density to a gamma distribution. We found that a gamma surface fits the posterior quite well, and further that five points are enough to provide a good estimate of its surface. This remains true even for neutral scenarios, where the log-likelihood functions are generally flatter (**fig. S1**).

3.2 Impact of E&R experimental design on detecting targets of selection

Bait-ER not only models the evolution of allele frequency trajectories but it also considers aspects of the experimental design specific to E&R studies. Bait-ER can thus be used to gauge the impact of particular experimental conditions in pinpointing targets of selection. We simulated allele frequency trajectories by considering a range of experimental parameters, including the number and span of sampled time points, the number of replicated populations, and the coverage. Each of these settings was tested in different population scenarios that we defined by varying the population size, starting allele frequency, and selection coefficient. We assessed the error of the estimated selection coefficients by calculating the absolute bias in relation to the true value. In total, we investigated 576 scenarios (**Supplementary table S2**). The heatmaps in **figure 3A-C** show the error for each scenario.

The heatmaps A, B, and C of **figure 3** show that the initial frequency is a determining factor in the accuracy of $\hat{\sigma}$ in E&R experiments. We observed that trajectories starting at very low frequencies (around 0.01) may provide unreliable estimates of σ. However, $\hat{\sigma}$'s accuracy on those trajectories can be improved by either increasing the sequencing depth or the number of replicates. Similar results have

been obtained using other methods such as in Kofler and Schlötterer (2014) and Taus et al. (2017). Designs with high coverage and several replicates may be appropriate when potential selective loci appear at low frequencies (e.g., dilution experiments). Surprisingly, alternative sampling schemes do not seem to substantially impact the accuracy of σ (**supplementary text S1**). These results have practical importance because sampling additional time points is time-consuming and significantly increases the cost of E&R experiments.

3.2.1 A note on population size

When using Bait-ER to estimate selection coefficients, one needs to specify the effective population size N_e. However, as effective population size and strength of selection are intertwined, mispecifying N_e will directly affect estimates of selection. The effective population size is often not known at the start of the experiment, but plenty of methods can estimate it from genomic data, e.g., Jonas et al. (2016). To assess the impact of mispecifying the effective population size on $p(\sigma)$, we simulated allele frequency trajectories using a fixed population size of 300 individuals. We then ran Bait-ER setting the effective population size to 100 or 1000. By doing so, we are increasing and decreasing, respectively, the strength of genetic drift relative to the true simulated population.

Bait-ER seems to produce highly accurate estimates of σ regardless of varying the effective population size (**fig. 4** and **supplementary fig. S5**). This is the case since misspecifying N_e merely rescales time in terms of Moran events rather than changing the relationship between N_e and the number of Moran events in the process. Further, we observed that the BFs are generally higher when the specified N_e is greater than the true value, suggesting that an increased false positive rate. The opposite pattern is observed when the population size one specifies is lower than the real parameter. Additionally, we investigated the relationship between BFs computed with the true N_e and those produced under a misspecified N_e. We found that these BFs are highly correlated (Spearman's correlation coefficients were always higher than 0.99; **fig. 4** and **supplementary fig. S5**). Taken together, our results indicate one should use a more stringent BF acceptance threshold if estimates of the real effective population size have wide confidence intervals.

Furthermore, we assessed Bait-ER's computational performance by comparing the relative CPU time while varying several user-defined experimental parameters. We found that increasing the effective population affects our software's computational performance most substantially (31-fold increase in CPU time when increasing the simulated population size from 300 to 1000 individuals; supplementary **table S1**).

3.3 Benchmarking Bait-ER with LLS, CLEAR and WFABC

3.3.1 Simulated Moran trajectories

To compare the performance of Bait-ER to that of other relevant software, we set out to simulate Moran frequency trajectories under the base experiment conditions described above. We tested Bait-ER as well as CLEAR (Iranmehr et al., 2017), LLS (Taus et al., 2017) and WFABC (Foll et al., 2015) on 100 trajectories for 4 starting frequencies (ranging from 1% to 50%) and 4 selection coefficients ($0 \leqslant N_e\sigma \leqslant 10$). All population parameters were tested for both 75 and 150 generations of experimental evolution. **Figure 5** shows the σ estimates for all methods under two starting frequency scenarios – 10% and 50%. CLEAR and LLS largely agree with Bait-ER's estimates of σ, even though the level of statistical significance is often not the same. It is evident that LLS produces estimates that are not as accurate as CLEAR's. This might have to do with the former not explicitly considering sampling bias in pool-seq data as a direct source of error. On the other hand, WFABC systematically disagrees with Bait-ER's estimates because its distribution is very skewed towards high selection coefficients. This is perhaps unsurprising given that WFABC does not consider replicate populations nor does it account for finite sequencing depth unlike the other three methods. We have included WFABC in our study for comparing Bait-ER with another Bayesian method. However, WFABC was not designed for E&R experiments, hence its poor performance on our simulated datasets.

Despite Bait-ER and CLEAR generally agreeing on the estimates of σ, there are a few trajectories that result in divergent estimates between these two methods. This is the case for those estimates seen on the top right plot in the leftmost corner in **figure 5**. In these few cases, Bait-ER overestimates σ, whereas CLEAR produces estimates closer to the true value. We investigated these further, and it seems that most alleles get lost – or fixed if we consider the alternative allele – between the first and the second time point. Furthermore, Bait-ER's BFs are also significant (mean approx. 6.4), suggesting our method is reflecting consistency across replicates (for further details, see **fig. S6**).

Regarding computational performance, Bait-ER seems to be the fastest of the four methods, even though it is comparable to WFABC (see **fig. 6**). However, we tested WFABC on the first replicate population data rather than the five experimental replicates used for the three remaining methods. Additionally, WFABC does not provide any statistical testing output such as a Bayes Factor. All these features make Bait-ER more thorough and just as fast as WFABC. In contrast, CLEAR and LLS are slower than the other two approaches. While CLEAR takes less than 40 seconds on average to analyse 100 sites, LLS is the slowest of the four, averaging around 4 minutes. Overall, these results suggest Bait-ER is just as accurate and potentially faster than other currently available approaches, which makes it a good resource for testing and inferring selection from genome-wide polymorphism datasets.

3.3.2 Complex simulation scenarios with recombination

For a more comprehensive study of Bait-ER's performance, we have analysed a complex simulated dataset produced by Vlachos et al. (2019). The authors simulated an E&R experiment inspired by the experimental setup of Barghi et al. (2019) and used polymorphism data from a *Drosophila melanogaster* population. In particular, we choose to focus on the classic sweep scenario, which is one of three complex scenarios simulated in Vlachos et al. (2019). For the sweep scenario, each experiment had 30 targets of selection randomly distributed along the chromosome arm. Each SNP was simulated with a fixed selection coefficient of 0.05. This dataset is key for benchmarking software like Bait-ER because it accounts for varying rates of recombination along the genome as well as replicated populations.

ROC (Receiver Operating Characteristic) curves are compared for five methods, Bait-ER, CLEAR, the CMH test (Agresti, 2003), LLS and WFABC, similarly to figure 2A in Vlachos et al. (2019). Bait-ER performs well with an average true positive rate of 80% at a 0.2% false positive rate (**fig. 7**). Its performance is as good as the CMH test's, but it does underperform slightly in comparison to CLEAR. Bait-ER, CLEAR and the CMH test greatly overperform LLS and WFABC.

ROC curves serve the purpose of showing how a method's level of statistical significance compares to other methods', may it be a p-value or a BF. It addresses whether the method places the true targets of selection amongst its highest scoring hits. While this is informative, it fails to account for the importance of finding a suitable significance threshold. For example, **figure 7** suggests that Bait-ER and the CHM test perform very similarly. However, the CHM test returns more potential targets than Bait-ER when comparable thresholds are used for both methods (e.g. **figure 10** that shows the comparison between Bait-ER logBFs and CMH test p-values for a real *D. simulans* dataset). This indicates that Bait-ER is more conservative and that the CMH test is more prone to producing false positives.

To assess why Bait-ER seems to be outperformed by CLEAR, we further investigated CLEAR's selection coefficient estimates. Comparison of selection coefficients estimated by Bait-ER and CLEAR showed that Bait-ER is slightly more accurate at estimating true targets' σ (**fig. S7**). In addition, it seems that those trajectories that scored highest with CLEAR are also the highest Bait-ER $\hat{\sigma}$ (**fig. S8**). Overall, Bait-ER and CLEAR perform to a similar high standard. However, the frequency variance filter implemented in Bait-ER seems to explain our method's slight underperformance shown in **figure 7**. Whilst the two method's false positive rates seem to be comparable, Bait-ER excluded a few selected sites from further analyses as they had changed very little in frequency throughout the experiment. Despite having excluded fewer than 70 (out of 30 targets times 100 experiments) targets of selection, Bait-ER's filtering step has also classified approximately the same amount of neutral trajectories for being too flat for inferring selection.

Overall, our results indicate that the dataset might be impacted by a phenomenon known as Hill-Robertson Interference (HRI) (Hill and Robertson, 1966). This is when adaptation is hindered by linked positively selected loci. It can result in incomplete sweeps, which are often hard to detect. Bait-ER estimated scaled selection coefficients ranged from 5.85 to 43.2, which suggests each target was under strong selection. Such values should be enough for selection to overcome genetic drift unless there is some degree of interference between selected sites within a 16Mb region. For the undetected targets of selection, the HRI effect and inconsistent responses between replicate populations might cause Bait-ER not to perform optimally.

3.4 Analysing E&R data from hot adapted Drosophila simulans populations

We have applied Bait-ER to a real E&R dataset that was published by Barghi et al. (2019). The authors exposed 10 experimental replicates of a *Drosophila simulans* population to a new temperature regime for 60 generations. Each replicate was surveyed using pool-seq every 10 generations. This dataset is particularly suited to demonstrate the relevance of our method, as Barghi et al. (2019) observed a strikingly heterogeneous response across the 10 replicates. The highly polygenic basis of adaptation has proved challenging to measure and summarise thus far.

The *D. simulans* genome dataset is composed of six genomic elements: chromosomes 2-4 and chromosome X. For each element, we have estimated selection parameters using Bait-ER. **Figure 8** shows a Manhattan plot of BFs for the right arm of chromosome 3. We can observe that there are two distinct peaks across the chromosome arm that seem highly significant (BF greater than 9). These two peaks – one at the start and another just before the centre of the chromosome – should correspond to loci that responded strongly to selection in the new lab environment. Such regions display a consistent increase in frequency across replicate populations. Overall, there are only a few other peaks that exhibit very strong evidence for selection across the genome (**fig. S10**). Those are located on chromosomes 2L, 2R and 3L. When compared to the CMH test results as per Barghi et al., Bait-ER's most prominent peaks seem to largely agree with those produced by the CMH (see **fig. S11**). The same is true for high BF regions on chromosomes 2L and 2R where there are similarly located p-value chimneys at the start of these genomic elements (**fig. S12**). Both Bait-ER and the CMH test did not produce clear signals of selection on chromosomes 3L, 4 and on the X.

One of the advantages of Bait-ER is that we have implemented a Bayesian approach for estimating selection parameters, which means we can calculate both the mean and variance of the posterior distributions. To examine both of these statistics, we looked into how the posterior variance varies as a function of mean σ. **Figure 9** shows the relationship between variance and mean selection coefficient for the X chromosome. We observe that the highest mean values also correspond to those with the

highest variance. Interestingly, most of those do seem to be statistically significant at a fairly lenient threshold (BF = 2). This suggests that the strongest response to selection, i.e. the highest estimated σ values, are also those showing a highly heterogeneous response across replicates. The remaining genomic elements seem to show similar patterns, apart from chromosome 4 (see **fig. S13**). This is consistent with other reports that inferring selection on this chromosome is rather difficult due to its size and low levels of polymorphism (Jensen et al., 2002).

Finally, we compared the p-values obtained by Barghi et al. (2019) and the BFs computed by Bait-ER. Barghi and colleagues performed genome-wide testing for targets of selection between first and last time points using the CMH test. The tests seem to largely agree for the most significant BFs correspond to the most significant p-values. However, Bait-ER appears to be more conservative than the CMH test. This follows from the finding that there is quite a substantial proportion of loci (less than 10% of all loci) that are deemed significant by a p-value threshold of 0.01, which are not accepted as such by Bait-ER. This is true even for a BF threshold of 2 such as that shown in **figure 10** for chromosome 2L.

Overall, Bait-ER performs well on such small experimental population. Bait-ER was designed to account for strong genetic drift on such small populations, hence the use of a discrete-population state space. It is rather conservative and produces only a few very significant peaks across the genome, which suggests it has a low false positive rate. Most of the genome produced BFs greater than 2, indicating that there is not enough resolution to narrow down candidate regions to specific genes despite those very significant peaks. Barghi et al. (2019) argue that there is strong genetic redundancy caused by a highly polygenic response to selection in their experiment. Despite Bait-ER modelling sweep-like scenarios rather than the evolution of a quantitative trait, the overall elevated BF signal across the genome might indicate that the genetic basis of adaptation to this new temperature regime is rather polygenic.

3.5 Conclusions and future directions

One of the main aims of E&R studies is to find targets of selection in genome-wide datasets. For that, we developed an approach that uses time series allele frequency data to estimate selection parameters. Bait-ER is a flexible statistical approach for inferring and testing for selection in laboratory experiments. It is faster and just as accurate as other relevant software. In addition, Bait-ER's implementation of the Moran model makes it suitable for experimental set-ups with overlapping generations. This is an advantage since it can be used in many large scale E&R designs.

Our results suggest that Bait-ER's inference is mostly affected by low starting allele frequencies. However, this can be overcome should the sequencing depth or the number of experimental replicates be increased. Although increasing the number of replicates might increase the cost of setting up an E&R experiment quite substantially, an improved sequencing depth is certainly within reach. This

interesting result might help guide future research. Encouragingly, Bait-ER performed well at small manageable population sizes, suggesting replication is key, but large populations are not necessarily required for achieving good results.

One aspect of time series polymorphism datasets that is worth our attention is that of missing data. It is sometimes the case that there is no frequency data at consecutive time points for a given trajectory. In the future, we will extend Bait-ER to allow for missing time points. Such a feature will enable one not to discard alleles for which not all time points have been sequenced. Using a probabilistic approach to estimate missing allele frequencies, Bait-ER will then be able to cope with missing data and estimate selection parameters.

Our approach assumes that each site in the genome is independent from one another. However, all selected sites will affect neighbouring loci. Neutral sites will increase in frequency towards fixation along which the true targets of selection. This causes allele frequencies to co-vary and selection to be overestimated around selected sites. On the other hand, the interference between positively selected sites, the HRI effect, leads to underestimation of the selection coefficients. Bait-ER can be extended to explicitly account for linkage, which decays with distance from any given locus under selection. Modelling the evolution of linked sites is not trivial, but it can be achieved if one includes information on the recombination landscape in the future.

4 Material and Methods

4.1 Modelling allele trajectories

Let us assume that there is a biallelic locus with two alleles, A and a. The evolution of allele A in time is fully characterised by a frequency trajectory in the state space $\{n_A, (N-n)_a\}$, where n is the total number of individuals that carry allele A (in a population of size N). Supposing the allele evolves according to the Moran model, the transition rates for the process are the following

$$\begin{aligned} n \to n-1 &: \ \tfrac{n(N-n)}{N} \\ n \to n+1 &: \ \tfrac{n(N-n)}{N}(1+\sigma) \end{aligned} \quad , \tag{1}$$

where $1+\sigma$ is the fitness of any A-type offspring and σ the selection coefficient for allele A. If $\sigma = 0$, i.e. A is evolving neutrally, then none of the alleles is preferred at reproduction. Let X_t be the number of copies of A in a population of N individuals; the probability of a given allele trajectory $\boldsymbol{X}$ can be defined using the Markov property as

$$p(\boldsymbol{X} \mid \sigma) = p(X_0 = x_0) \prod_{t=1}^{T} p(X_t = x_t \mid X_{t-1} = x_{t-1}, \sigma) \quad , \tag{2}$$

where T is the total number of time points measured in generations at which the trajectory was assayed. The conditional probability on the left-hand side of the equation has one calculating $X_t = e^{Q d_t} X_{t-1}$, where Q is the rate matrix defined in (1) and d_t the difference in number of generations between time point t and $t-1$. The probability of a single allele frequency trajectory can be generalised for R replicates by assuming their independence

$$p(\boldsymbol{X} \mid \sigma) = \prod_{r=1}^{R} p(X_0^r = x_0^r) \prod_{t=1}^{T} p(X_t^r = x_t^r \mid X_{t-1}^r = x_{t-1}^r, \sigma) \quad . \tag{3}$$

The main caveat for pool-seq data is the fact that it provides estimates for allele frequencies, not true frequencies. For that reason, we assume that the allele counts are generated by a binomial sampling process $B(n/N, C)$ which depends on the frequency of allele A and the total sequencing depth C obtained by pool-seq. We then recalculate the probability of the Moran states given an observed allele count c by considering the inverse of the binomial sampling process

$$p(\{nA, (N-n)a\} \mid \{c, C\}) \propto \binom{C}{c} \left(\frac{n}{N}\right)^{c} \left(1 - \frac{n}{N}\right)^{C-c}, \quad n = 0, \ldots, N \quad . \tag{4}$$

This step is key for it corrects for sampling noise generated during data acquisition. This is particularly relevant for low frequency alleles and poorly covered loci.

4.2 Inferential framework

We used a Bayesian framework to estimate σ. It requires allele counts and coverage for each time point and replicate population $\{\boldsymbol{c}, \boldsymbol{C}\}$ at each position as input. The posterior distribution can then be obtained by

$$p(\sigma|\{\boldsymbol{c}, \boldsymbol{C}\}) \propto p(\sigma)p(\{\boldsymbol{c}, \boldsymbol{C}\}|\sigma) \quad . \tag{5}$$

Our algorithm is defined using a gamma prior on σ. The posterior cannot be formally obtained, hence we define a grid of σ values for which we calculate the posterior density. Estimating the posterior distribution $p(\sigma|\{\boldsymbol{c}, \boldsymbol{C}\})$ is a time consuming part of our algorithm because the likelihood is computationally costly to compute. To avoid this burden, we fit the posterior to a gamma density

$$\log p(\sigma|\{\boldsymbol{c}, \boldsymbol{C}\}) = c + (\alpha - 1)\log\sigma - \beta\sigma \quad , \tag{6}$$

where α and β are the shape and rate parameters, respectively, and c the normalization constant. The gamma fitting represents a good trade-off between complexity, since it only requires two parameters, but its density may take many shapes. As one requires the values of α and β that best fit the gamma density for further analyses, we find the least squares estimates of α and β (and c), such that the error is minimal. The estimation is as follows

$$\begin{aligned}\hat{\alpha} &= \frac{-(s_2 s_4 + s_4^2 - s_6 - s_7)(s_1^2 - s_8) - (s_3 + s_1 s_2 + s_1 s_4 + s_5)(s_1 s_4 - s_5)}{s_7 s_1^2 - 2 s_4 s_5 s_1 + s_5^2 + s_4^2 s_8 - s_7 s_8} \quad \wedge \\ \hat{\beta} &= \frac{-s_3 s_4^2 + s_2 s_5 s_4 + s_1 s_6 s_4 - s_5 s_6 - s_1 s_2 s_7 + s_3 s_7}{s_7 s_1^2 - 2 s_4 s_5 s_1 + s_5^2 + s_4^2 s_8 - s_7 s_8} \quad ,\end{aligned} \tag{7}$$

where $s_1 = \sum_i x_i/N$, $s_2 = \sum_i y_i/N$, $s_3 = \sum_i x_i y_i/N$, $s_4 = \sum_i \log x_i/N$, $s_5 = \sum_i x_i \log x_i/N$, $s_6 = \sum_i y_i \log x_i/N$, $s_7 = \sum_i \log^2 x_i/N$ and $s_8 = \sum_i x_i^2/N$. We evaluated the fitting of the gamma density for neutral and selected loci, and observed that a gamma surface with five points describes the log posterior of selected and neutral loci quite suitably (**fig. S1**).

Additionally, Bait-ER was implemented with an allele frequency variance filter that is applied before performing the inferential step of our algorithm. This filtering process excludes any trajectories that do not vary or vary very little throughout the experiment from further analyses. To do that, we assess the trajectories' frequency increments and exclude loci with frequency variance lower than 0.01. These correspond to cases where trajectories are too flat to perform any statistical inference on. Trajectories such as these typically have both inflated $\hat{\sigma}$ and BFs.

Bait-ER is implemented in C++ and freely available for download at `https://github.com/mrborges23/Bait-ER` (accessed on December 12th 2020). Here, we provide a tutorial on how to compile and run Bait-ER, including a toy example with 100 loci taken from Barghi

et al. (2019).

4.3 Simulated data

We tested our algorithm's performance under several biologically relevant scenarios using (1) a Moran model allele frequency trajectory simulator, and (2) the individual-based forward simulation software MimicrEE2 (Vlachos and Kofler, 2018).

The Moran model simulator was used, firstly, for benchmarking Bait-ER's performance across a range of experimental conditions, and, secondly, to compare our estimates of σ to those of CLEAR (Iranmehr et al., 2017), LLS (Taus et al., 2017) and WFABC (Foll et al., 2015). We started out by testing Bait-ER under different combinations of experimental and population parameters. A full description of these parameters can be found in **table S2**. Scenarios that explored several experimental designs included those with varying coverage (20x, 60x and 100x), number of replicate populations (2, 5 and 10) and number of sampled time points (2, 5 and 11). In addition to simulating even sampling throughout the experiment, we tested our method on trajectories where we varied sampling towards the start or towards the end of said experiment. Total study length might also affect Bait-ER's estimation, therefore we tracked allele frequency trajectories for $0.2N_e$ and $0.4N_e$ generations.

We set out to compare Bait-ER to other selection estimation software using experimental parameters that resemble realistic E&R designs. Our base experiment replicate populations consist of 300 individuals that were sequenced to 300x coverage. There are five such replicates that were evenly sampled five times throughout the experiment. We then simulated 100 allele frequency trajectories for all starting frequencies and selection coefficients mentioned above. We simulated trajectories for $0.25N_e$ as well as $0.5N_e$ generations.

The performance of both CLEAR and LLS was assessed by running the software on any sync file with a fixed population size of 300 individuals (flag –N=300 and estimateSH(..., Ne = 300), respectively). Additionally, to estimate the selection coefficient under the LLS model, we used the estimateSH(...) function assuming allele codominance (argument h = 0.5). WFABC was tested with a fixed population size of N_e individuals (flag -n 300), lower and upper limit on the selection coefficient of -1 and 1, respectively (flags -min_s -1 and -max_s 1), maximum number of simulations of 10000 (flag -max_sims 10000) and four parallel processes (flag -n_threads 4). The program was run for 1200 seconds, after which the process timed out to prevent it from running indefinitely in case fails to converge. This caused trajectories with starting allele frequencies of 5% and 1% not to be analysed at all. We have thus only been able to include results for alleles starting at 10% and 50% frequencies.

We used data simulated by Vlachos et al. (2019) using MimicrEE2 (Vlachos and Kofler, 2018) to

benchmark Bait-ER and compare it with other relevant statistical methods. MimicrEE2 allows for whole chromosomes to be simulated under a wide range of parameters mimicking the effects of an E&R setup on allele frequencies. We used this data for it allows for testing our method including relevant biological parameters such as variation in recombination rate. Moreover, the simulated data were used to test the performance of other relevant statistical methods.This dataset consisted of 10 replicate experimental populations, and each experimental population consisted of 1,000 diploid organisms evolving for 60 generations. The haplotypes used to found the simulated populations were based on 2L chromosome polymorphism patterns from *Drosophila melanogaster* fly populations (Bastide et al., 2013). Recombination rate variation was based on the *D. melanogaster* recombination landscape (Comeron et al., 2012). Low recombination regions were removed from the dataset. 30 segregating loci were randomly picked to be targets of selection with a selection coefficient of 0.05. Sites were initially segregating at a frequency between 0.05 and 0.95.

4.4 Application

We applied our algorithm to the recently published dataset from an E&R experiment in 10 replicates of a *Drosophila simulans* population to a hot temperature regime for 60 generations (Barghi et al., 2019). All populations were kept at a census size of 1000 individuals. The experimental regime consisted of light and temperature varying every 12 hours. The temperature was set at either 18°C or 28°C to mimic night and day, respectively. The authors extracted genomic DNA from each replicate population every 10 generations using pool-seq. The polymorphism datasets are available at `https://doi.org/10.5061/dryad.rr137kn` in sync format. The full dataset consists of more than 5 million SNPs. We subsampled the data such that Bait-ER was tested on 20% of the SNPs. Subsampling was performed randomly across the whole genome.

5 Acknowledgements

This work was supported by the Vienna Science and Technology Fund (WWTF) through project MA16-064. CK received funding from the Royal Society (RG170315) and Carnegie Trust (RIG007474). The computational results presented have been partly achieved using the St Andrews Bioinformatics Unit (StABU), which is funded by a Wellcome Trust ISSF award (grant 105621/Z/14/Z). We are grateful to Peter Thorpe for his help with using the StABU cluster. We thank Neda Barghi, Abigail Laver and Mike Ritchie for helpful discussions and suggestions on an early version of Bait-ER. The genome-wide scan for selected loci in the Barghi et al. (2019) dataset was conducted using the Vienna Scientific Cluster (VSC).

6 Tables

	WFABC	**CLEAR**	**LLS**	**Bait-ER**
Inference approach	Approximate Bayesian computation	Maximum likelihood	Linear least squares*	Bayesian
Hypothesis testing	• Bayes factors • Depends heavily on summary statistics	• Likelihood-ratio tests • Empirical p-values based on genome-wide drift simulations	• Empirical simulated p-values based on simulations of allele trajectories	• Bayes factors • Based on the posterior distribution
Assumptions	• WF model	• WF model	• WF and Moran model • The allele frequencies vary linearly with the selection coefficients • Weak selection	• Time-continuous Moran model
Accounts for replicates	No	Yes	Yes	Yes
Accounts for sequencing noise	No	Yes	No	Yes
Reference	(Foll et al. 2015)	(Iranmehr et al. 2017)	(Taus et al. 2017)	This study

Table 1: **Currently available software for estimating selection coefficients in E&R experiments.**[a]

[a]The table describes several features of each method namely: i) the approach used for inferring selection coefficients, ii) whether it performs hypothesis testing or not, iii) what sort of assumptions are made about the underlying population genetics model, iv) its overall computational and inference performance, v) whether it accounts for multiple replicate populations, and vi) whether it accounts for sampling variance due to sequencing noise. WF: Wright-Fisher. *LLS under the assumption of linearity is equivalent to a maximum likelihood approach.

7 Figures

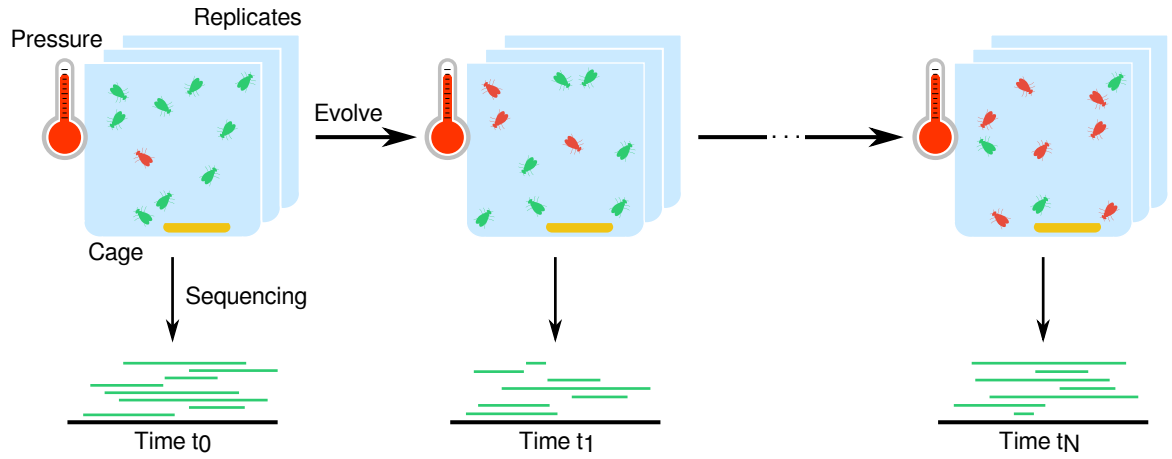

Figure 1: **Example of an E&R experimental setup.** E&R experiments expose several replicated populations (e.g., of flies, yeast, viruses) to a selective pressure (e.g., temperature, food regimes) for a specific number of generations t_T. The replicated populations are surveyed at several time points by whole-genome sequencing, which allows one to quantify changes in allele frequencies over time.

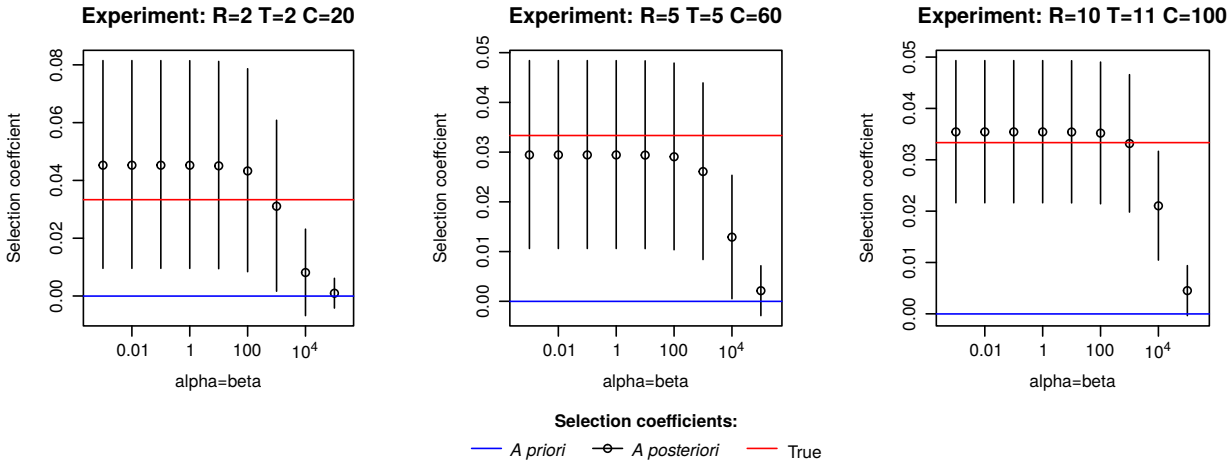

Figure 2: **Impact of the prior on the posterior estimates of the selection coefficients.** The posterior distribution of σ was calculated using gamma priors $G(\alpha, \beta)$, where α and β are the shape and rate parameters. We set $\alpha = \beta$ and allowed β to vary from 0.001 to 10^5 (i.e. ranging from a very uninformative to a very informative prior, respectively). The different priors were tested under three E&R experiment scenarios: the first was a sparse experimental design (coverage (C) = 20x, number of time points (T) = 2 and number of replicates (R) = 2), while the second mimicked a standard set up (C = 60x, T = 5 and R = 5). Finally, the third scenario had the most thorough experimental conditions (C = 100x, T = 11 and R = 10). Red lines indicate the true value of σ. Blue lines point to the mean of the prior imposed on σ. Black lines and points correspond to the posterior mean of σ and credibility intervals at 0.95.

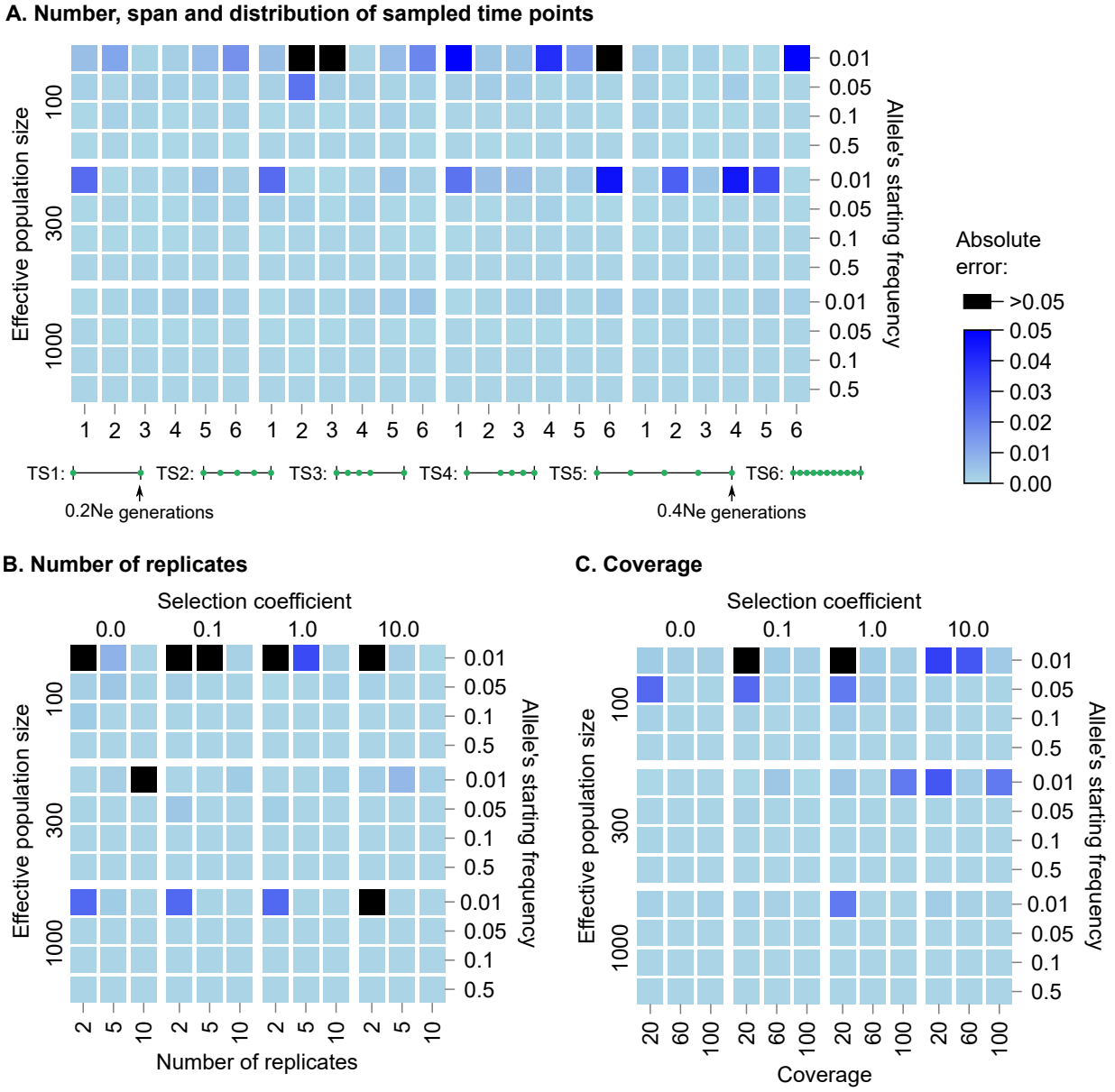

Figure 3: **Impact of E&R experimental design on the estimated selection coefficients.** Each square of the heatmap represents the error of the estimated selection coefficients, i.e., the absolute difference between the estimated and the true σ: $|\hat{\sigma} - \sigma|$, for a range of population dynamics and E&R experimental conditions. **(A)** Number, span and distribution of sampled time points. The six time schemes differ according to the following criteria: most time schemes have five sampling events, except for TS1 and TS6, which have two and eleven time points, respectively; all time schemes have a total span of $N_e/5$ generations, except for TS5, which has double the span ($2N_e/5$); uniform sampling was used in most scenarios but for TS3, which is more heavily sampled during the first half of the experiment, and TS4, during the second half. The two maximum experiment lengths considered ($0.2N_e$ and $0.4N_e$) were chosen based on typical E&R experimental designs. **(B)** number of replicates. **(C)** coverage. To test the experimental conditions, we defined a base experiment with five replicates, five uniformly distributed time points (total span of $0.20Ne$ generations) and a coverage of 60x. The complete set of results is shown in **supplementary fig. S2-S5**.

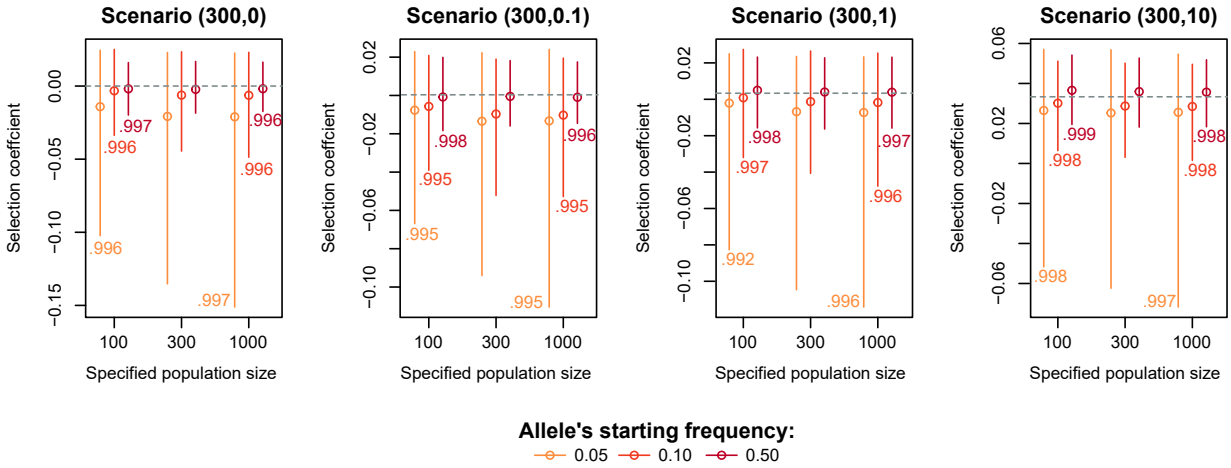

Figure 4: **Impact of the user-specified population size on the estimation of selection coefficients.** The plots show the distribution of the estimated selection coefficients where the population size is misspecified. Vertical lines and points indicate the interquartile range and median selection coefficient. Each plot represents a specific scenario that was simulated by varying the population size, the true selection coefficient (indicated within brackets ($N_e, N_e\sigma$)) and starting allele frequency (indicated by the yellow-to-red colour gradient). The numbers next to each bar correspond to the Spearman's correlation coefficient, which correlates the BFs of the 100 replicated trajectories between the cases where we have either under- and overspecified the population size ($N_e = 100$ or 1000, respectively) and the case where we use the true population size ($N_e = 300$). Regarding simulated experimental design, we defined a base experiment with five replicates, five uniformly distributed time points (total span of $0.20Ne$ generations) and a coverage of 60x.

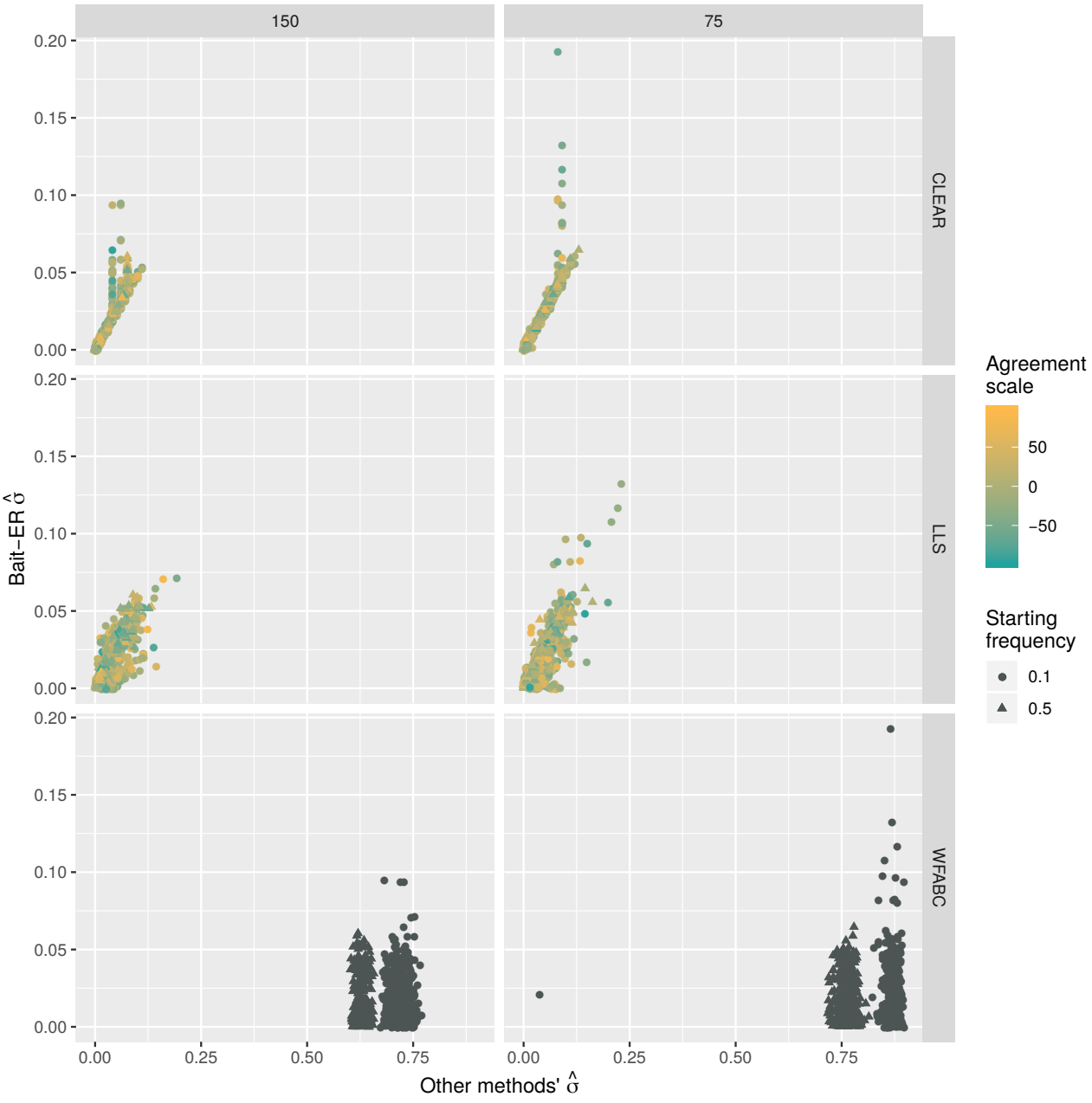

Figure 5: **Comparison of estimates of σ produced by Bait-ER versus CLEAR, LLS and WFABC.** These plots include estimates for those trajectories simulated with starting frequencies of 10% and 50%. Each horizontal panel compares Bait-ER's estimates to those produced by CLEAR, LLS and WFABC from top to bottom row, respectively. The left and right hand side panels correspond to two different experiment lengths: 150 and 75 generations, respectively. LLS returned NA's for 4 out of 800 trajectories which were excluded from these graphs.

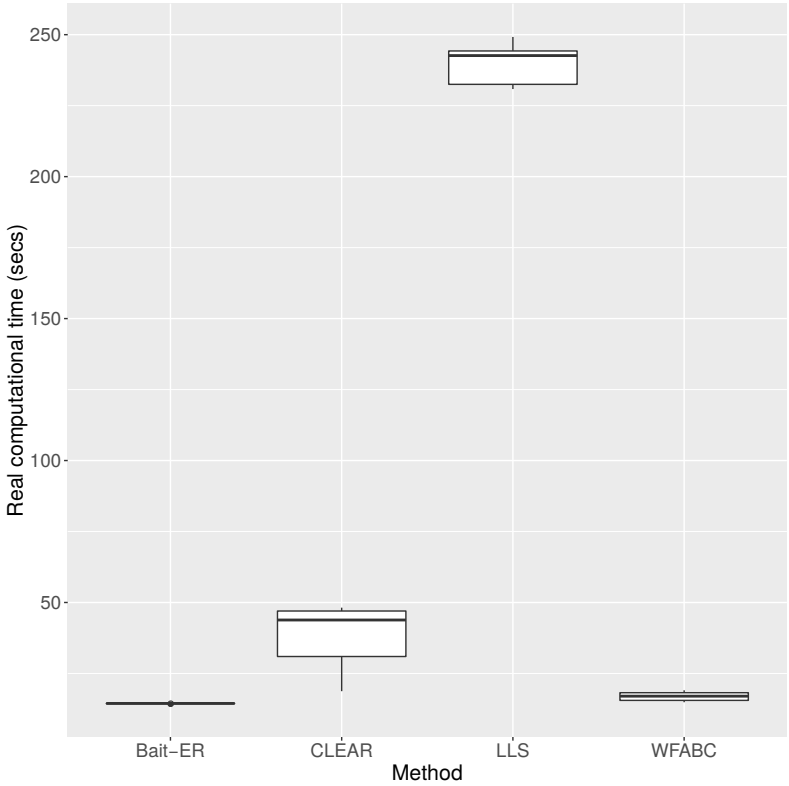

Figure 6: **Real computational time for Bait-ER and the other three approaches tested.** From left to right, computational time in seconds including both inference and hypothesis testing for Bait-ER, CLEAR, LLS and WFABC is shown here. Similarly to figure 5, these boxplots include estimates for those trajectories simulated with starting frequencies of 10% and 50%, as well as both study lengths investigated, i.e. 150 and 75 generations. Those 4 NA's produced by LLS were again removed from these plots.

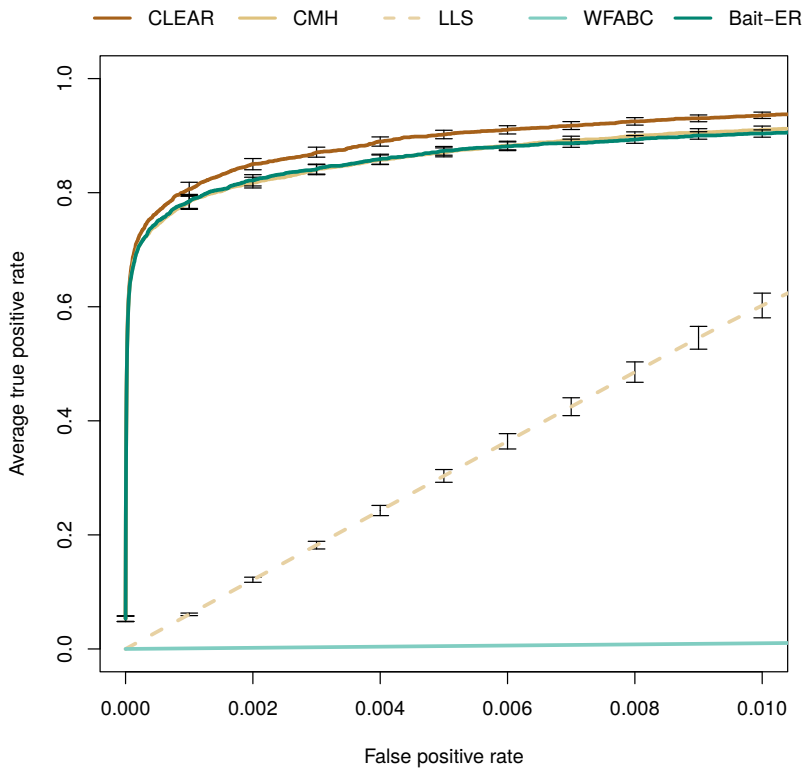

Figure 7: **Performance of Bait-ER and other software at testing for selection in a complex simulated dataset.** ROC (Receiver Operating Characteristic) curves for Bait-ER, CLEAR, CMH, LLS and WFABC under the classic sweep scenario simulated by Vlachos et al. (2019).

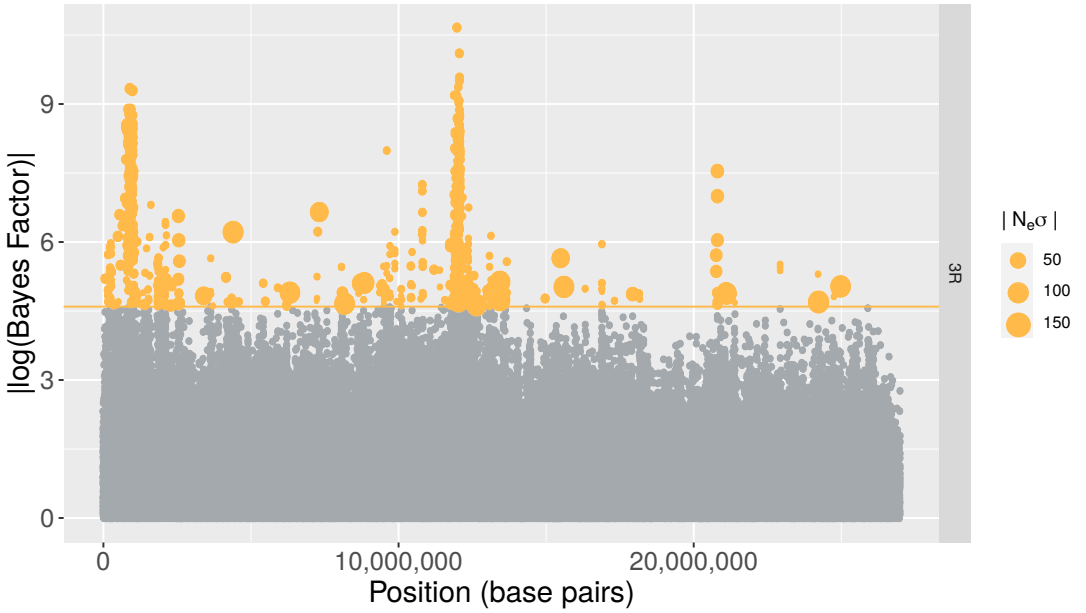

Figure 8: **Bayes Factors on chromosome 3R.** This Manhattan plot shows log-transformed Bayes Factors computed by Bait-ER for loci along the right arm of the 3rd chromosome in the Barghi et al. (2019) time series dataset. The orange line indicates a conservative threshold of approximately 4.6, which corresponds to log(0.99/0.01), meaning all points in orange have very strong evidence for these to be under selection. The SNPs that are significant at this level are sorted by size according to how strong Bait-ER's selection coefficients are. In other words, points are sized according to how strong the large selection coefficient is estimated to be.

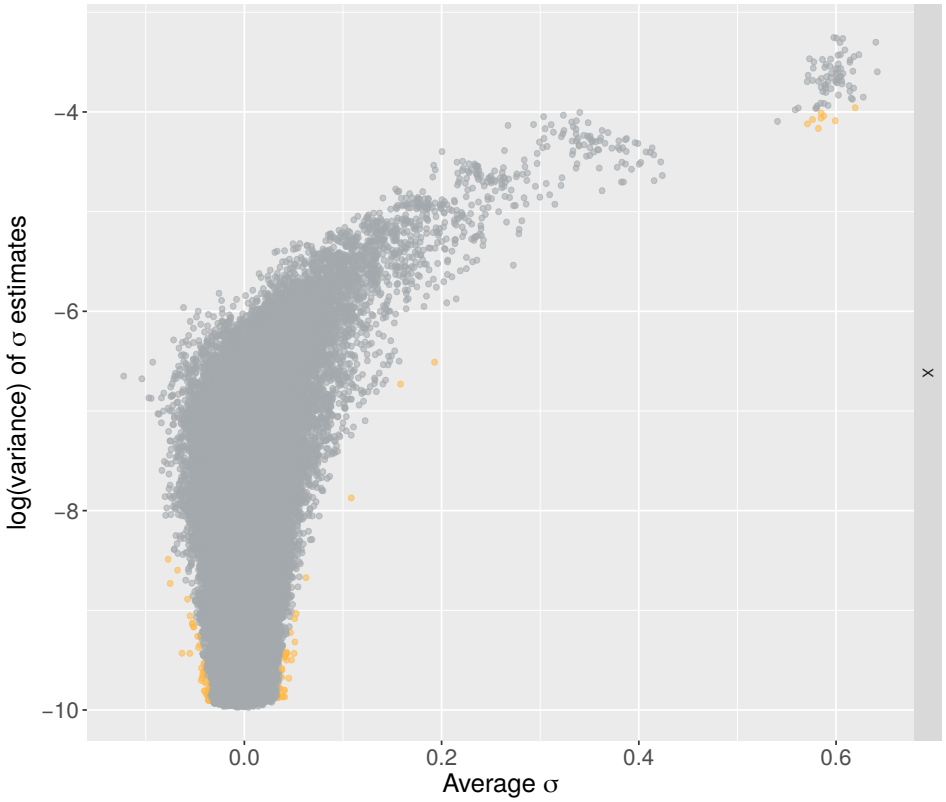

Figure 9: **Variance versus mean sigma on the X chromosome.** This graph compares log transformed variances in σ estimates to average σs. The variance is calculated using the inferred rate and shape parameters for the beta distribution, and the average σ is the mean value of the posterior distribution estimated by Bait-ER. Orange coloured points are significant at a conservative BF threshold of log(0.99/0.01), approx. 4.6.

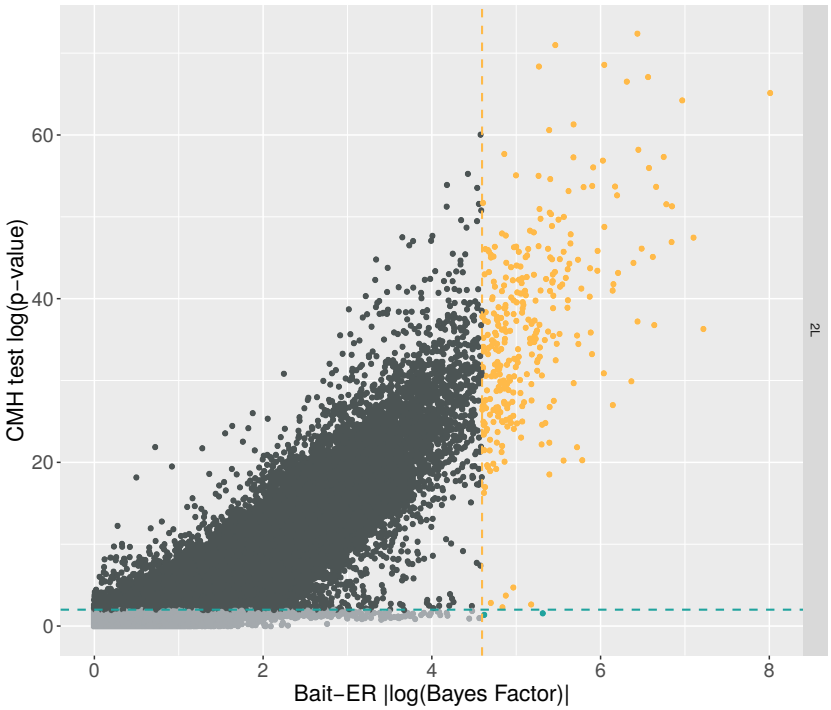

Figure 10: **Bait-ER's Bayes Factors versus CMH test's p-values on chromosome 2L.** Orange coloured points correspond to BFs which are greater than log(0.99/0.01) (approx. 4.6) and p-values less than or equal to 0.01, i.e. those that are considered significant by both tests. Blue coloured points indicated that the computed BF is greater than our threshold, but not significant according to the CMH test. Additionally, dark grey points are significant according to the CMH test, but not to Bait-ER, and light grey points are inferred not significant by both tests.

References

Agresti, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons.

Barghi, N., Tobler, R., Nolte, V., Jakšić, A. M., Mallard, F., Otte, K. A., Dolezal, M., Taus, T., Kofler, R., and Schlötterer, C. (2019). Genetic redundancy fuels polygenic adaptation in Drosophila. *PLoS Biol*, 17(2):e3000128.

Bastide, H., Betancourt, A., Nolte, V., Tobler, R., Stöbe, P., Futschik, A., and Schlötterer, C. (2013). A genome-wide, fine-scale map of natural pigmentation variation in drosophila melanogaster. *PLoS genetics*, 9(6):e1003534.

Burke, M. K., Liti, G., and Long, A. D. (2014). Standing Genetic Variation Drives Repeatable Experimental Evolution in Outcrossing Populations of Saccharomyces cerevisiae. *Molecular Biology and Evolution*, 31(12):3228–3239.

Cochran, W. G. (1954). Some Methods for Strengthening the Common χ 2 Tests. *Biometrics*, 10(4):417.

Comeron, J. M., Ratnappan, R., and Bailin, S. (2012). The many landscapes of recombination in drosophila melanogaster. *PLoS genetics*, 8(10):e1002905.

Debelle, A., Courtiol, A., Ritchie, M. G., and Snook, R. R. (2017). Mate choice intensifies motor signalling in Drosophila. *Animal Behaviour*, 133:169–187.

Foll, M., Shim, H., and Jensen, J. D. (2015). WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Molecular Ecology Resources*, 15(1):87–98.

Futschik, A. and Schlötterer, C. (2010). The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics*, 186(1):207–218.

Godwin, J. L., Vasudeva, R., Michalczyk, Ł., Martin, O. Y., Lumley, A. J., Chapman, T., and Gage, M. J. G. (2017). Experimental evolution reveals that sperm competition intensity selects for longer, more costly sperm. *Evolution Letters*, 1(2):102–113.

Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical Research*, 8(03):269.

Iranmehr, A., Akbari, A., Schlötterer, C., and Bafna, V. (2017). Clear: Composition of Likelihoods for Evolve and Resequence Experiments. *Genetics*, 206(2):1011–1023.

Jensen, M. A., Charlesworth, B., and Kreitman, M. (2002). Patterns of genetic variation at a chromosome 4 locus of Drosophila melanogaster and D. simulans. *Genetics*, 160(2):493–507.

Jonas, A., Taus, T., Kosiol, C., Schlotterer, C., and Futschik, A. (2016). Estimating the Effective Population Size from Temporal Allele Frequency Changes in Experimental Evolution. *Genetics*, 204(2):723–735.

Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., and Whitlock, M. C. (2012). Experimental evolution. *Trends in Ecology & Evolution*, 27(10):547–560.

Kofler, R., Pandey, R. V., and Schlötterer, C. (2011). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, 27(24):3435–3436.

488 Kofler, R. and Schlötterer, C. (2014). A Guide for the Design of Evolve and Resequencing Studies.
489 *Molecular Biology and Evolution*, 31(2):474–483.

490 Papkou, A., Guzella, T., Yang, W., Koepper, S., Pees, B., Schalkowski, R., Barg, M. C., Rosenstiel,
491 P. C., Teotónio, H., and Schulenburg, H. (2019). The genomic basis of red queen dynamics during
492 rapid reciprocal hostpathogen coevolution. *Proceedings of the National Academy of Sciences of the*
493 *United States of America*.

494 Schlötterer, C., Kofler, R., Versace, E., Tobler, R., and Franssen, S. U. (2015). Combining
495 experimental evolution with next-generation sequencing: a powerful tool to study adaptation from
496 standing genetic variation. *Heredity*, 114(5):431–440.

497 Taus, T., Futschik, A., and Schlötterer, C. (2017). Quantifying Selection with Pool-Seq Time Series
498 Data. *Molecular Biology and Evolution*, 34(11):3023–3034.

499 Topa, H., Jónás, Á., Kofler, R., Kosiol, C., and Honkela, A. (2015). Gaussian process test for
500 high-throughput sequencing time series: Application to experimental evolution. *Bioinformatics*,
501 31(11):1762–1770.

502 Turner, T. L., Stewart, A. D., Fields, A. T., Rice, W. R., and Tarone, A. M. (2011). Population-based
503 resequencing of experimentally evolved populations reveals the genetic basis of body size variation in
504 Drosophila melanogaster. *PLoS Genetics*, 7(3).

505 Vlachos, C., Burny, C., Pelizzola, M., Borges, R., Futschik, A., Kofler, R., and Schlötterer, C. (2019).
506 Benchmarking software tools for detecting and quantifying selection in evolve and resequencing
507 studies. *Genome Biology*, 20(1):1–11.

508 Vlachos, C. and Kofler, R. (2018). MimicrEE2: Genome-wide forward simulations of Evolve and
509 Resequencing studies. *PLOS Computational Biology*, 14(8):e1006413.

510 Wiberg, R. A. W., Gaggiotti, O. E., Morrissey, M. B., and Ritchie, M. G. (2017). Identifying consistent
511 allele frequency differences in studies of stratified populations. *Methods in Ecology and Evolution*,
512 2017(February):1–11.