Journal of Computational Mathematics and Data Science 10 (2024) 100091

Contents lists available at ScienceDirect

Journal of Computational Mathematics and Data Science

journal homepage: www.elsevier.com/locate/jcmds

Approximate Bayesian computational methods to estimate the strength of divergent selection in population genomics models

Martyna Lukaszewicz [a,c,d,*], Ousseini Issaka Salia [a,b,c,d,e], Paul A. Hohenlohe [a,b,c,d], Erkan O. Buzbas [a,b,c]

[a] Institute for Interdisciplinary Data Sciences (IIDS), University of Idaho, Moscow, ID, United States of America
[b] Institute for Modeling Collaboration and Innovation (IMCI), University of Idaho, Moscow, ID, United States of America
[c] Department of Mathematics and Statistical Science, University of Idaho, Moscow, ID, United States of America
[d] Department of Biological Sciences, University of Idaho, Moscow, ID, United States of America
[e] Department of Horticulture, Washington State University, Pullman, WA, United States of America

ARTICLE INFO

Keywords:
ABC
Mechanistic model
Kernel
Joint parameters estimation
Population genomics
Summary statistics

ABSTRACT

Statistical estimation of parameters in large models of evolutionary processes is often too computationally inefficient to pursue using exact model likelihoods, even with single-nucleotide polymorphism (SNP) data, which offers a way to reduce the size of genetic data while retaining relevant information. Approximate Bayesian Computation (ABC) to perform statistical inference about parameters of large models takes the advantage of simulations to bypass direct evaluation of model likelihoods. We develop a mechanistic model to simulate forward-in-time divergent selection with variable migration rates, modes of reproduction (sexual, asexual), length and number of migration-selection cycles. We investigate the computational feasibility of ABC to perform statistical inference and study the quality of estimates on the position of loci under selection and the strength of selection. To expand the parameter space of positions under selection, we enhance the model by implementing an outlier scan on summarized observed data. We evaluate the usefulness of summary statistics well-known to capture the strength of selection, and assess their informativeness under divergent selection. We also evaluate the effect of genetic drift with respect to an idealized deterministic model with single-locus selection. We discuss the role of the recombination rate as a confounding factor in estimating the strength of divergent selection, and emphasize its importance in break down of linkage disequilibrium (LD). We answer the question for which part of the parameter space of the model we recover strong signal for estimating the selection, and determine whether population differentiation-based summary statistics or LD–based summary statistics perform well in estimating selection.

1. Introduction

Divergent selection occurs when populations adapt to contrasting environments, causing the accumulation of genomic differences due to differential selective pressure in these environments. Identity of loci under divergent selection and estimating the strength of divergent selection at these loci play a key role in detecting divergent selection, which can be a driving force of speciation [1]. Here, we aim to build theoretical and experimental models of divergent selection and assess the computational feasibility and quality of statistical inference under these models. Specifically we investigate whether the loci under divergent selection can be identified, and the strength of selection can be estimated with reasonable precision using state of the art simulation-based statistical methods.

* Corresponding author at: Institute for Interdisciplinary Data Sciences (IIDS), University of Idaho, Moscow, ID, United States of America.
E-mail address: martyna.lukaszewicz@gmail.com (M. Lukaszewicz).

https://doi.org/10.1016/j.jcmds.2024.100091
Received 18 December 2023; Received in revised form 15 January 2024; Accepted 31 January 2024

Our experimental system is the baker's yeast (*Saccharomyces cerevisiae*) as a fast-evolving model organism. Yeast is an eukaryotic model organism with a moderate number of linear chromosomes (n=16) and a genome-wide number of cross-over events per meiosis that is comparable to larger eukaryotes [2], but with small genome size (12 Mb). Yeast can undergo both asexual and sexual reproduction. Crossing of yeast strains and meiosis (sporulation) can be experimentally controlled. The sporulation when sexual reproduction takes place requires starvation of yeast from nitrogen, glucose, and carbon source [3]. Single haploid spores can be isolated and sequenced to determine haplotype phase across the genome as well as precisely map meiotic crossover, gene conversion events, and recombination rate heterogeneity across the genome [2].

Some of the challenges in detecting divergent selection are as follows. On the statistical side, divergent selection models that we investigate are the result of stochastic processes on genomes through generations and therefore, they are large models. This causes two main challenges for statistical inference. The first is formulating a workable exact likelihood function. Simulation-based computational methods working with approximate model likelihoods as opposed to exact likelihoods partially solves this problem. The second is making inference scalable for large simulation studies to investigate model properties so that we know what to expect when we perform inference using data. A well-known simulation-based method based on approximate likelihoods is Approximate Bayesian Computation (ABC). ABC uses simulated data under an assumed model using a large set of parameter values to generate an approximate sample from the posterior distribution of interest by bypassing the evaluation of the likelihood function [4].

On the genetics side, physical linkage of nearby neutral loci to the locus under selection may lead to genetic hitchhiking [5] (i.e., the change in allele frequency at neutral loci), which contributes to the formation of genomic islands [6], thereby making identifying the locus under selection challenging. The magnitude of the hitchhiking effect is modeled to be proportional to the distance of the actual locus under selection, influenced by the recombination rate [7], for which we apply an average, fixed genomic recombination rate.

Keeping these two main difficulties in perspective, we design a computationally efficient simulator scalable at genomic scales to study the behavior of a divergent selection model. We use our simulator in a large simulation study to assess which statistics are informative about divergent selection, under variable migration, mode of reproduction (sexual or asexual), and in the presence of fixed recombination. We also investigate the bias and variance of estimators in estimating the strength of selection.

To address these, in Section 2, we describe the experimental design of the biological yeast system and model parameters controlled in laboratory settings, the developed theoretical population genetics model, and how the experimental design of the theoretical model translates to the biological model. Next, in Section 3, we describe the constraints of the process of generation of the data, define the model parameters, and we build the simulator for the theoretical population genetics model. In Section 4 we describe the ABC, and how it assesses the summary statistics from the simulator output data in the estimation of the model parameters, with and without the initial outlier scan on the observed data where the outlier scan reduces the parameter space of potential positions under selection to be considered. In Section 5 we present our results on estimating model parameters from both with and without the loci outlier scan, and for four different scenarios under which the initial parameters vary for the outlier scan method. To conclude, in Section 6 we discuss the application of the simulation study combined with the ABC to bypass the likelihood function, the role of the parameter space of selection and for which part we recover a strong signal, what summary statistics out of the tested with our model outperform the others, and finally the confounding role of the recombination rate to make inferences about the biological yeast system.

2. Model

In this section, we first describe the biological system in which *yeast* is used as a model organism in an experimentally controlled environment of selective pressures, type of reproduction, and strength of migration. Then we describe a theoretical population genetics model as an idealization of this biological system.

2.1. Description of the biological system and experimental design

We have genetically engineered an obligate diploid cross between our two focal yeast strains, a North American oak isolate (YPS128) and an Australian vineyard isolate (DBVPG1106). The YPS128 and DBVPG1106 strains form the biological system experimental setup and they differ by over 70,000 SNPs, an average of 1 SNP per ~165bp [8]. We treat each SNP as a *locus*. The two strains form a diploid pool, without recombination, followed by sporulation of randomly selected two parents, during which recombination occurs, to create an offspring ancestral pool at time $t = 0$. Half of the ancestral becomes a founding population assigned to evolve in sodium dodecyl sulfate (SDS) and the other half of the ancestral pool becomes a founding population assigned to evolve in sodium chloride (NaCl). These environments induce differential selective environmental pressures in two populations.

The biological experiment was performed under four different migration-selection cycles treatments, with three replicates per treatment. The four different treatments were as follows: no migration and no sporulation (asexual reproduction), no migration and sporulation (sexual reproduction), 20% migration and sporulation, and 50% migration and sporulation. During the divergence with gene flow, the populations evolve asexually in one of the mediums for 5 days, which is assumed to be equivalent to $t^* = 50$ generations. Then, one of the four scenarios of migration conditional on the type of sexual reproduction is implemented for one generation. This cycle is repeated four times, resulting in evolution under divergent selection for approximately 200 generations.

2.2. Description of the theoretical population genetics model

We assume diploid organisms that differ by L bi-allelic loci. Each population is of constant effective population size N_e, and they evolve in non-overlapping generations. We obtain the populations $\boldsymbol{X_0}$ and $\boldsymbol{Y_0}$ at time $t = 0$ by crossing the founding populations at all loci consisting of all private alleles at L loci, then incorporating the recombination events into the process, and finally we evolve the populations through generations as follows.

We assume that the two parental genomes for an offspring are uniformly randomly (independently of each other) distributed, one parent from each population. The number of recombinations on the offspring's genome, n_r, is binomially distributed with probability r. We define the position vector on which these n_r events happen by $\boldsymbol{L_r}$, $\|\boldsymbol{L_r}\| \leq L$, at which recombination events are uniformly randomly (independently of each other) distributed on L loci. The two parental genomes are recombined in positions defined by $\boldsymbol{L_r}$ to obtain the offspring genome.

The reproduction in recurring cycles of t^* generations: We start from $t = 0$, isolate $\boldsymbol{X_t}$ and $\boldsymbol{Y_t}$ from each other and implement asexual reproduction for a sequence of $t^* - 1$ generations. These $t^* - 1$ generations allow for population and loci-specific divergent selection to act on each population.

We let s_i to denote the selection coefficient at locus $i \in \{1, 2, \ldots, \|\boldsymbol{L_s}\|\}$, where the fitness of the reference allele a_i under selection is $(1 + s_i)$ if population carrier has a_i copy, else $(1 + 0)$. We assume that selection effects are additive across loci on the genome such that the fitness of an individual at time $t - 1$, in population j ($j \in \{\boldsymbol{X}, \boldsymbol{Y}\}$) (**Part A** of Fig. 1) is

$$\omega_n^{(j,t-1)} = \sum_{i=1}^{\|\boldsymbol{L_s}\|} \left(1 + \mathbf{I}_{n\{a_i \in (j,t-1)\}} s_i\right). \tag{1}$$

Then, for each distinct genome, the probability of including an offspring at generation t is multinomially distributed with the probability of successes proportional to their normalized fitnesses given by

$$p_n = \frac{\omega_n^{(j,t-1)}}{\sum_{n=1}^{N_e} \omega_n^{(j,t-1)}}. \tag{2}$$

From generation $t^* - 1$ to t^* populations undergo asexual or sexual reproduction with recombination and symmetric migration. Migration rate from $\boldsymbol{X_{t^*-1}}$ to $\boldsymbol{Y_{t^*}}$ and from $\boldsymbol{Y_{t^*-1}}$ to $\boldsymbol{X_{t^*}}$ is denoted by m. We sample uniformly randomly (independently of each other) $N_e m$ parents to migrate from population j to the other population. After migration, reproduction is either sexual ($sex = 1$), or asexual ($sex = 0$). If $sex = 0$, an offspring is an exact copy of a single parent chosen uniformly randomly. If $sex = 1$, we choose two parents uniformly randomly (independently of each other) from the same population. The recombination steps are the same as described above (**Part B** of Fig. 1).

The next reproduction cycle starts at generation $t^* + 1$ (**Part C** of Fig. 1), for a total of specified number of cycles n_{cycles}, with final generation occurring before migration, i.e. $t_{final} = n_{cycles} t^* - 1$. Visual representation of the experimental design of recombination rates and modes of reproduction is seen in Fig. 2

3. Model parametrization, the data, and the divergent selection simulator

In this section, we delineate the constraints on the data-generating process, define the parameters of the population genetic model, and finally build a simulator. The divergent selection simulator allows us to explore the divergent selection model motivated by the yeast populations built in the lab and it constitutes one of our main contributions.

3.1. Model parametrization and the data

To capture differences in population $\boldsymbol{X}$ relative to population $\boldsymbol{Y}$, we economically use signed selection coefficients. We arbitrarily fix the reference allele a_i at locus i with selection coefficient s_i ($i \in \{1, 2, \ldots, \|\boldsymbol{L_s}\|\}$) for the other allele at that locus in population $\boldsymbol{X}$. A negative s_i means that for allele a_i at locus i, an individual in population $\boldsymbol{X}$ has a lower fitness in comparison to an individual in population $\boldsymbol{Y}$. Unlike in population $\boldsymbol{X}$, the fitness of the reference allele a_i under selection in population $\boldsymbol{X}$ is always $(1 + 0)$ for a carrier in population $\boldsymbol{Y}$, regardless of whether the carrier in population $\boldsymbol{Y}$ has the a_i copy or not.

SNPs in yeast occur for approximately one in every 165bp. We assume equally spaced SNPs in the genome and rescale the genome-wide recombination rate proportionally. We further assume that there is at most one crossover event between consecutive SNPs. We let $SNP_{spacing}$ be the spacings between L SNPs, which allows us to denote the total genome length by $L \times SNP_{spacing}$. The recombination rate is a function of L and $SNP_{spacing}$ is a constant: $r = f(L)$, $SNP_{spacing} = C$.

In the experimental laboratory setup, the migration rate and the length of migration-selection cycles t^* are controlled, and the recombination rate can be estimated as described earlier in Section 1, and in Section 2.1. The migration rates are fixed at $\{0, 0.2, 0.5\}$. For the recombination rate r, we use values from the literature and assume them fixed and known. A typical computational procedure mimicking sequenced yeast genome informed by our laboratory procedures is as follows. We picked to simulate $L = 1,500$ SNPs because it is computationally scalable. This number translates into about one-third on mean of SNPs of a chromosome that the two yeast strains in our biological model system from Section 2.1 differ by over 70,000 SNPs [8]. The mean recombination rate of *Saccharomyces cerevisiae* has been estimated as 3.5×10^{-6} Morgans/bp in the literature [9], with inferred genome-wide recombination profiles from sequenced isolates from an advanced intercross line (AIL) to be as high as 3.0×10^{-5} Morgans/bp for a two-way cross at genome hotspots [10]. Here, for our laboratory procedures mimicking the sequenced yeast, we fixed the genome recombination

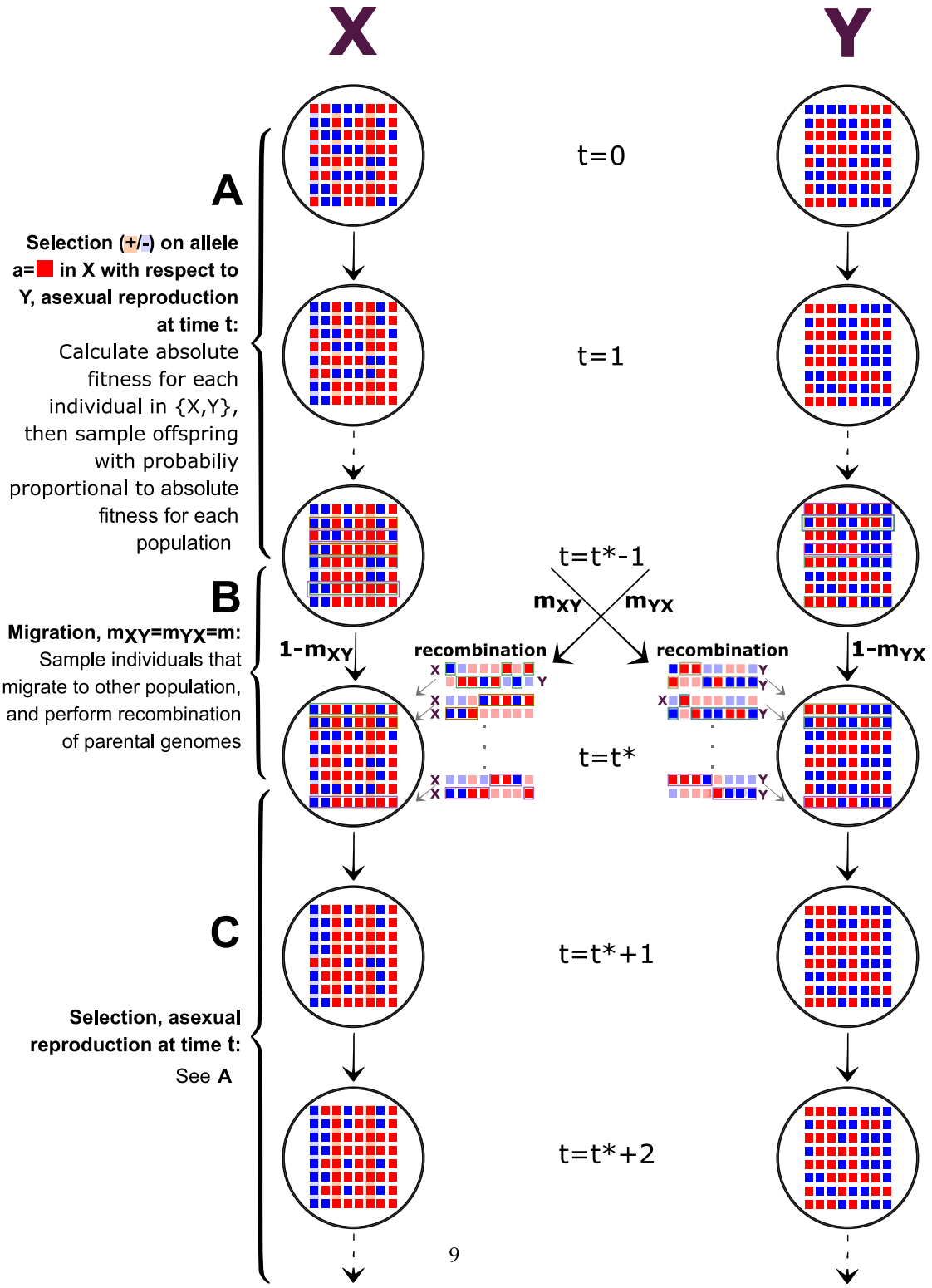

Fig. 1. Population genetic model for divergent selection: A. Population divergence for $t^* - 1$ generations during which reproduction is asexual and the absolute fitness depends on the allele-specific copy under which selection acts upon. Each offspring is genetically identical to its only parent and an individual is chosen to be a parent with probability proportional to its fitness. B. Symmetric migration rates m_{XY} and m_{YX} between $t^* - 1$ and t^* generation. Neutral evolution with recombination. C. Second population divergence.

rate of $r = 2.0 \times 10^{-5}$ Morgans/bp. One way to think about this value is as a best statistical estimate for recombination hotspots in yeast. For the other scenarios with a smaller number of SNPs for computational time efficiency, we scaled the recombination rate to $r = 3.0 \times 10^{-4}$, proportionally to the expected number of recombination events.

Fixing unknown parameter values to their point estimates instead of jointly estimating them is not ideal. Recombination rate is known to be a particularly problematic parameter in population genetic models with selection and recombination. This is due to the fact that the two evolutionary processes might generate statistically similar signals in genetics. In models the recombination rate r was varied in our simulations the signal from divergent selection was confounded to the degree that there were no useful statistical inferences to distinguish true loci under selection. However, our work is a first serious effort to model divergent selection and to explore the statistical properties of estimates of a parameter of interest (strength of divergent selection) in a large evolutionary model. Free model parameters of not direct interest such as recombination rate cause statistical identifiability issues that jeopardize the statistical inference. These issues have not been solved in population genetics (except in small models) and their discussion is beyond the scope of this paper. Statistical identifiability of model parameters kicks in when there is a large number of interacting parameters in a population genetics model that are not fixed, but vary.

Table 1 describes the full parametrization with assumed values for fixed and known parameters and prior distributions for unknown and to be estimated parameters. We tried to choose reasonable support for the prior distributions based on yeast literature [10–15]. To be explicit in probability functions of the model representation, in addition to standard conditionality notation separating the observables and parameters, we denote the fixed and known parameters by $\boldsymbol{K} = (r, t^*, n_{cycles}, N_e, L, SNP_{spacing})$ where r is the recombination rate per genome per generation, t^* is the number of generations between reproduction cycles, n_{cyclec} is the number of reproduction cycles, N_e is the effective population size, and L is the number of SNPs, $SNP_{spacing}$ is the spacing on the genome between each SNP. We write the joint probability mass function generating the data as

$$P(\boldsymbol{x}|\boldsymbol{s}, m, sex, \boldsymbol{l}; \boldsymbol{K}) \tag{3}$$

where, $\boldsymbol{x}$ is an N_e by L matrix of zeros and ones for the SNP data for each population, $\boldsymbol{s} = (s_1, s_2, \ldots, s_{\|\boldsymbol{L_s}\|})$ is the vector of signed selection coefficients at each locus, m is the symmetric migration rate per generation between the two populations, sex is the mode of reproduction, $\boldsymbol{l} = (l_1, l_2, \ldots, l_{\|\boldsymbol{L_s}\|})$ is the vector of indicators to denote loci under selection.

3.2. Simulator for the data generating process

Below we describe how the data are generated with the simulator. We input effective population size N_e, number of SNPs per carrier to be simulated L, loci corresponding to SNPs under selection $\boldsymbol{L_s}$, alleles under selection a, recombination rate r, selection coefficients at loci under selection s, migration rate between populations m, number of generations between migration-selection cycles t^*, type of reproduction during migration generations sex, spacing of SNPs on the genome, assumed equal spacing $SNP_{spacing}$, and number total number of generations to simulate t_{final}.

The model parameters that are estimated from the output simulator are the loci under selection and corresponding selection coefficients, although we allow the migration rate between populations m and the type of reproduction during migration generations sex to vary between simulations but we do not estimate them. The estimated model parameters are sampled from the joint prior distribution, i.e. $\theta_i^* \sim P(\boldsymbol{s}, m, sex, \boldsymbol{l}; \boldsymbol{K})$.

Our simulated data of bi-allelic SNPs are $N_e \times L$ matrices per population represented by $\boldsymbol{x}$.

The SNPs of each carrier in the population correspond to the matrix row. For each of N_e carries per population an allele a is sampled from discrete uniform distribution bound on $[0, 1]$ and replicated L times, corresponding to the probability of 0.5 of carriers from founding F_0 population $\boldsymbol{X}$ and of 0.5 from founding F_0 population $\boldsymbol{Y}$, to build F_1 populations $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively.

Offsprings are then generated from the recombination of two parental genomes per offspring. For 1 to N_e per population, the first parent is sampled with equal probability from $\boldsymbol{X_{F_1}}$, and the second parent is sampled with equal probability from $\boldsymbol{Y_{F_1}}$, both from a discrete uniform distribution, i.e. $p_1 \sim$ Discrete Unif$[1, N_e]$ from $\boldsymbol{X_{F_1}}$ and $p_2 \sim$ Discrete Unif$[1, N_e]$ from $\boldsymbol{Y_{F_1}}$. The number of recombination events n_r between p_1 and p_2 are sampled from $Bin(L \times SNP_{spacing} - 1, r)$, with r corresponding to genomic recombination rate and $SNP_{spacing}$ corresponding to SNP spacing on the genome. Loci of recombination events, $\boldsymbol{L_r}$, are sampled n_r times from $\boldsymbol{L_r} \sim$ Discrete Unif$[n_r, L \times SNP_{spacing} - 1]$. If a random variable rv sampled from Uniform distribution on $[0, 1]$ is less than 0.5, i.e. if $rv \sim$ Unif$[0, 1] < 0.5$, p_1 is reassigned to p_2 and p_2 is reassigned to p_1. The recombination of p_1 and p_2 then starts from the genome of p_1 and alternates between two parents to form an offspring. This concludes founding F_2 at time $t = 0$.

Then, the selection and asexual reproduction occur for $t^* - 1$ generations. Absolute fitness of parents $\omega_n^{(j,t-1)}$ for each population j in $\boldsymbol{X}, \boldsymbol{Y}$ for each carrier n is calculated. For loci under selection specified by an input vector $\boldsymbol{L_s}$, absolute fitness of parent at $t-1$ is $\omega_n^{(j,t-1)} = \sum_{i=1}^{\|\boldsymbol{L_s}\|} \left(1 + \mathbf{I}_{n\{a_i \in (j,t-1)\} s_i}\right)$, where s_i correspond to non-zero selection coefficients specified by vector $\boldsymbol{s}^* = (s_1^*, s_2^*, \ldots, s_{\|\boldsymbol{L_s}\|}^*)$, and alleles under selection specified by vector $\boldsymbol{a}^* = (a_1^*, a_2^*, \ldots, a_{\|\boldsymbol{L_s}\|}^*)$.

Probability of carrier in the population p_n having an offspring is the normalized absolute fitness $p_n = \frac{\omega_n^{(j,t-1)}}{\sum_{n=1}^{N_e} \omega_n^{(j,t-1)}}$. The N_e offspring are $N_e^{(j,t)} \sim$ Multinomial$(N_e, \boldsymbol{p})$, with normalized absolute fitness probabilities $\boldsymbol{p} = \left(p_1, p_2, \ldots, p_{N_e}\right)$.

At generation, t^* migration between populations takes place, specified by migration rate m, with the corresponding type of reproduction specified by $sex \in \{0, 1\}$.

The migration-selection cycles as described in **Part A** and **Part B** respectively are repeated for a total specified number of cycles n_{cycles}, with ending on the final generation before migration, i.e. $t_{final} = n_{cycles} t^* - 1$. The output of the data of the x matrices of dimension $N_e \times L$ per population is then ready for model inference by mapping the data to summary statistics and the ABC.

Table 1
Properties of parameters in the model.

Parameter	Status	Value	Prior	Motive
$SNP_{spacing}$	fixed and known	$\frac{\text{genome length}}{\text{number of SNPs strains differ by}} = 165$	—-	genome length of about 12 Mpb and strains differ by about 73,000 SNPs [8]
μ_{n_r}	fixed and known	4.95	—-	expected value of number of crossovers= $\mathbb{E}(\text{Bin}(L \times SNP_{spacing} - 1, r))$
r	fixed and known	2.0×10^{-5} Morgans/bp, and 3.0×10^{-4} Morgans/bp for $L = 1,500$ and $L = 100$ respectively	—-	already done in literature [10,14,15]
l_i	fixed, but unknown for 1st method; not fixed and unknown for 2nd method, where potentially identified within outlier scan, specified by set_{l_i}	—-	$set_{l_i} = \{l_{i=1} = L/3, l_{i=2} = L/2, l_{i=3} = 2L/3\}$ for 1st method; set_{l_i} identified by initial outlier scan of 5% for 2nd method	fixed loci as discrete value parameters at which selection acts to reduce the parameter space to $\{l_{i=1} = L/3, l_{i=2} = L/2, l_{i=3} = 2L/3\}$ and setting the values for rest of loci to zero for 1st method; less selection parameter values $s_i \mid I = i$ to estimate in ABC after initial outlier scan, and setting the values for loci not identified by outlier scan to zero for 2nd method
$s_i \mid I = i$	unknown and estimated	—-	Unif$[-0.25, 0.25]$ for $t^* = 5$, and Unif$[-0.025, 0.025]$ for $t^* = 50$ respectively	considered a large range of parameter space for $s_i \mid I = i$ in yeast [16], and 10-fold smaller for 10-fold longer selection generation cycles for comparison
n_r	random and unknown	—-	—-	number of recombination events sampled from Bin$(L \times SNP_{spacing} - 1, r)$
$\boldsymbol{L_s}$	fixed but unknown	—-	Discrete Unif$[set_{l_i}]$	vector $\boldsymbol{L_s}$ of length 0, 1, or 2 of loci under selection to reduce the parameter space
$\boldsymbol{L_r}$	random and unknown	—-	—-	vector $\boldsymbol{L_r}$ of length n_r of loci of recombination events sampled from Discrete Unif$[n_r, L \times SNP_{spacing} - 1]$
m	fixed but unknown	—-	Discrete Unif$[set_m]$, $set_m = \{0.0, 0.2, 0.5\}$	methods of estimation of m already developed [12] but considering some variability (set)
N_e	fixed and known	10,000	—-	already done in literature [13,17]
L	fixed and known	1,500 or 100	—-	1,500 SNPs is about one-third on average of SNPs of a chromosome for a yeast cross [8]; 100 SNPs is for faster computational time

(*continued on next page*)

Table 1 (*continued*).

t^*	fixed and known	5 or 50	—-	short enough selection generation cycles of $t^* = 5$ to consider large range of parameter space for s_i; longer selection generation cycles of $t^* = 50$ for comparison
n_{cycles}	fixed and known	2 or 4	—-	two different population mixing frequencies for comparison on the signature of selection
sex	fixed but unknown	—-	Discrete Unif$[set_{sex}]$, $set_{sex} = \{0, 1\}$	yeast system can reproduce sexually or asexually [11]

4. Inference about model parameters via approximate Bayesian computation (ABC)

The probability distribution given in expression (3) is not available in closed form and the joint likelihood of the data cannot be evaluated given the parameters. There are no exact methods to perform statistical inference about the unknown model parameters in this case. As a practicable solution to the problem of performing inference about the model parameters, we employ ABC to sample the posterior distribution of parameters [4,18–24]. ABC bypasses the explicit evaluation of the joint likelihood thereby making simulation-based inference feasible when model likelihoods cannot be evaluated. Statistical inference in ABC is characterized by two main approximations. The first approximation is due to substituting the exact likelihood of the data with a kernel-based numerical approximation. The second approximation is due to substituting the likelihood of the data with the likelihood of the summary statistics. Whether and by how much these approximations affect the quality of the inference depends on the size of the model generating the data and the computational budget available to increase accuracy of the approximation. For the first approximation, good practices have been established. Assessing the quality of the second approximation, however, is particularly challenging in a class of models where there are no known sufficient statistics for unknown parameters. Our divergent selection model falls into this class. In this section, we investigate the usefulness of some population differentiation statistics to perform inference about the parameters of our divergent selection model.

Based on expression (3) for the probability model generating the data, we denote the joint posterior distribution of parameters given the data $\boldsymbol{x}$ and fixed and known parameters as

$$P(\boldsymbol{s}, m, sex, \boldsymbol{l}|\boldsymbol{x}; \boldsymbol{K}) \propto P(\boldsymbol{x}|\boldsymbol{s}, m, sex, \boldsymbol{l}; \boldsymbol{K})P(\boldsymbol{s}, m, sex, \boldsymbol{l}; \boldsymbol{K}), \tag{4}$$

where $P(\boldsymbol{x}|\boldsymbol{s}, m, sex, \boldsymbol{l}; \boldsymbol{K})$ is the joint likelihood of the data and $P(\boldsymbol{s}, m, sex, \boldsymbol{l}; \boldsymbol{K})$ is the joint prior distribution of unknown parameters. Incorporating the two approximations described in the previous paragraph, we write the likelihood as

$$P(\boldsymbol{s}, m, sex, \boldsymbol{l}|\boldsymbol{SumStat}; \boldsymbol{K}) \propto P(\boldsymbol{SumStat}|\boldsymbol{s}, m, sex, \boldsymbol{l}; \boldsymbol{K})P(\boldsymbol{s}, m, sex, \boldsymbol{l}; \boldsymbol{K}), \tag{5}$$

where $\boldsymbol{SumStat}$ is summary statistics. In the next two subsections, we evaluate some useful summary statistics in the context of our model.

4.1. Summary statistics and outlier scan on summarized observed data

The effect of divergent selection on genomes between samples of two populations can be quantified by well-known statistics that measure of genetic differentiation. An example is Wright's fixation index [25], F_{ST}. For bi-allelic loci F_{ST} defined by:

$$F_{ST} = \sigma_p^2/[\bar{p}(1 - \bar{p})] \tag{6}$$

which measures allele frequency differentiation among the sampled populations [26]. Here, σ_p^2 is the variance in allele frequency among sampled populations, and $\bar{p}$ is the mean allele frequency in sampled populations. A signed version of F_{ST}, which we denote by $signF_{ST}$ obeys the following: **if** $(p_X - p_Y) < 0$, $signF_{ST} = -F_{ST}$, **else** $signF_{ST} = F_{ST}$, such that $signF_{ST} \in [-1, 1]$. This statistic captures the information about which sampled population is undergoing selection advantageous with respect to the other sampled population.

At the genomic scale, it is often computationally infeasible to approximate the likelihood based on F_{ST} jointly for all loci. A practicable remedy, which we follow here, is first to determine a set of candidate loci under selection manifesting only outlier values of F_{ST} and consider the likelihood based on these loci. We take the outlier cutoff to be F_{ST} outside of the 95% of all F_{ST} values in the data [27]. The selection coefficients can take on values within a range defined in Table 1, it is not a single value, for which we can derive an estimated F_{ST} cut-off value, therefore we do not consider just large F_{ST} values. In the simulator, the summary statistics outliers correspond to specific SNPs. Only the specific loci of SNPs that were detected by the outlier test are considered

as potentially under selection when simulating data sets for ABC. The rest of the loci (non-outliers) are fixed to as no selection, i.e. $s_i = 0$.

Sample F_{ST} is not a sufficient statistic for any parameter of a divergent selection model since large values of sample F_{ST} are not only the result of divergent selection. Sample F_{ST} from expression (6) measures genetic variation among sampled populations by assessment of variance between and within sampled populations by calculating allele frequency differences. The sample F_{ST} values have been found to be correlated with the recombination rate [28]. Recombination rate between loci and the number of generations of recombination influence LD decay with genetic distance. Under a neutral evolution model, genetic drift is the only driving force of changes in allele frequencies. Many generations will be required for a new variant to reach a high frequency, and the surrounding LD will decay due to recombination events [29–31].

In our paper we test an LD-based summary statistic – Cross Population Extended Haplotype Homozygosity ($XP-EHH$) [32,33] – along with the sample F_{ST}. In order to derive sample $XP-EHH$, one must first calculate Extended Haplotype Homozygosity (EHH) summary statistics for each population.

The EHH between two loci is defined as the probability that the number of distinct haplotypes G_v in a genomic region up to a distance v from the locus are equal to each other. For N_e carriers per population with possible alleles of either 0 or 1 per locus, with each group z, $z = 1, 2, \ldots, G_v$, having n_z haplotypes, EHH is:

$$EHH(v) = \frac{\sum_{z=1}^{G_v} \binom{n_z}{2}}{\binom{n_0}{2} + \binom{n_1}{2}} \tag{7}$$

[34]. At distance, $v = 0$ the $EHH(v = 0)$ for each locus with respect to itself is always 1 and the $EHH(v)$ values decay as the v increases, which is the decay of LD from each core haplotype [32,35]. The $XP-EHH$, compares the integrated extended haplotype homozygosity, EHH, between two populations [32,34]. Specifically, $XP-EHH$ is the ratio of the EHH between populations $\boldsymbol{X}$ and $\boldsymbol{Y}$ integrated over the genome. If recombination rates in the model were allowed to vary widely across the genome between and within populations, the EHH statistic can be interpreted as a measure of selection only after suitable normalization [34,36]. Our model assumes a mean recombination rate for simulated L SNPs, thus normalization is not required in our model.

Calculating $EHH(v)$ values for locus 1 to L, and corresponding distances v 0 to $L-1$, it yields an L by L symmetric matrix for each of the two populations. The $XP-EHH$ then for population $\boldsymbol{X}$ and $\boldsymbol{Y}$ combined for each locus is just a vector of length L, and is given by:

$$XP-EHH(v) = \log\left(\frac{\int_D EHH_{pop_X}(v)\mathrm{d}v}{\int_D EHH_{pop_Y}(v)\mathrm{d}v}\right). \tag{8}$$

The integration domain D is a cut-off threshold below which the EHH values are set to 0.0, for which we pick D of 0.05 [35,37]. The $XP-EHH$ has the advantage of detecting selection on alleles that near fixation in one population but not both, population $\boldsymbol{X}$ and $\boldsymbol{Y}$ [34]. This fits our model where the population $\boldsymbol{X}$ evolves under the positive or negative selection with respect to population $\boldsymbol{Y}$, with occasional gene flow between two populations during migration generations.

4.2. Assessment of summary statistics

Following the outlier scan on the summarized observed data set (F_{ST} statistic) outside of the 95% of all F_{ST} values as described in the previous subsection, only the outlier loci in the data sets simulations identified by the outlier scan are considered as potentially under selection, with selection coefficients at non-outlier loci fixed to 0, and only the summary statistics corresponding to loci identified by outlier scan are inputted to Algorithm 1 and Algorithm 2. We assess the performance of summary statistics in terms of how well they capture the signal of observed selection coefficients by using the summary statistics in the ABC, then calculating the standard deviation of posterior distributions of selection coefficients, and mean square errors (MSEs), variance and squared-bias. We compare performance of the following summary statics: $signF_{ST}$, $XP-EHH$, $p_X - p_Y$, and $signF_{ST}$ with $XP-EHH$. We plot SNPs vs. MSE, observed selection coefficients vs. MSE, SNPs vs. squared-bias, observed selection coefficients vs. squared-bias, SNPs vs. variance, observed selection coefficients vs. variance, for the four summary statistics combinations, and for the mode of reproduction and migration rate in Section 5.

In our model p_X and p_Y represent allele frequency per locus in sampled population $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively, such that $p_X - p_Y$ is the difference between the allele proportions per locus between the two populations, with the expected value of zero for the founding F_2 populations, i.e. at generation $t = 0$. Additionally to testing the performance of summary statistics from the outlier scan on summarized observed data described above, we assessed the effect of genetic drift and found that the population size is large enough for genetic drift not to be an issue as the mean of 100,000 data sets converge to the deterministic model, as verified by simulations. Due to the convergence of the mean of 100,000 data sets and the expected value from the deterministic model, we determined most informative summary statistics performed of simulator output data ($t = t_{final}$) on the deterministic model based on the MSEs. We derived a summary statistic called $signF_{ST}$: **if** $(p_X - p_Y) < 0$, $signF_{ST} = -F_{ST}$, **else** $signF_{ST} = F_{ST}$. Due to the strong fit of superimposed plots of simulations and deterministic values [38], we evaluated $signF_{ST}$, F_{ST}, and $(p_X - p_Y)$ based on deterministic single locus model. We performed 1000 ABC tests with a tolerance rate of 0.1%. For each of the ABC iterations, a single simulated data set with known parameters was randomly drawn from the 100,000 data sets and assumed as observed data set. For each, the top 0.1% of data sets with the smallest Euclidean distances between observed and simulated summary statistics were accepted. On average of the 1000 ABC iterations, the lowest error was achieved with $signF_{ST}$ ($MSE = 5.50 \times 10^{-8}$). We compared ABC results to empirical results from the simulator from 100,000 simulations and achieved on average $MSE = 2.00 \times 10^{-3}$ with $signF_{ST}$ and

$MSE = 1.10 \times 10^{-2}$ with $(p_X - p_Y)$. We also examined parameter space via plots of $s_{L/2}$ vs. summary statistics and the relations are more often 1:1 with $signF_{ST}$ than with $(p_X - p_Y)$. In Section 5 we address $XP - EHH$ and $signF_{ST}$ and expand it to a much larger scope of parameter space.

4.3. Inference about model parameters

We estimate the free model parameters in $P(\boldsymbol{s}, m, sex, \boldsymbol{l}|\boldsymbol{SumStats}; \boldsymbol{K})$ given by expression (5), using an ABC-rejection algorithm (Algorithm 1) and the ABC with the linear regression adjustment (Algorithm 2). The term $\boldsymbol{x}$ represents, the data from our bi-allelic SNPs of individuals in two populations. The simulated SNPs are $N_e \times L$ matrices per population represented by $\boldsymbol{x}$, and $\boldsymbol{x}$ summarized by summary statistics $SumStat$ by mapping $\boldsymbol{SumStat} = S(\boldsymbol{x})$. The $nsim$ observations denoted by $\boldsymbol{x}$ generated from the model are independent and identically distributed, i.i.d. The $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X}$ is the space in which the data sits.

ABC makes two approximations, the first one is the mapping of $\boldsymbol{x}$ to $\boldsymbol{SumStat}$, and the second is accepting summary statistics $\boldsymbol{SumStat}$ within a tolerance rate from the observed $\boldsymbol{SumStat_{obs}}$. The simulations are a mechanistic process that involves random sampling, which can be thought of as an influence of stochastic processes such as genetic drift. The ABC facilitates in model parameters estimation by accepting parameters corresponding to summarized data sets within a tolerance rate that is partly due to the stochastic effect. The ABC outputs a posterior distribution of accepted parameters $P(\boldsymbol{s}, m, sex, \boldsymbol{l}|\boldsymbol{SumStat_m}; \boldsymbol{K})$, $m \sim$ Discrete Unif$[set_m]$, where $set_m \in [0, 1]$. In the ABC-rejection (Algorithm 1) we calculate the Euclidean distance d_i, $i = 1, 2, \ldots, nsim$ technique for each of the simulated and summarized by the summary statistics data set [39,40] to scale summary statistics across the $nstat$ dimensional space. The summary statistics are standardized by the median absolute deviation SD, j, $j = 1, 2, \ldots, nstat$, such that each of the $nstat$ summary statistics per dimension approximately equally contribute to the ABC analysis. Further details on the Euclidean distance are described in Appendix A.

Algorithm 1 ABC-rejection algorithm for summary statistics calculated from data sets simulated from the **Simulator** and their corresponding parameters from the prior distribution. Input: proportion of simulations to accept, number of simulated data sets, number of summary statistics per simulation.

1: Input: tolerance rate, number of simulations, number of summary statistics.
2: Input parameter values. ▷ Parameter values sampled from the prior (Table 1).
3: Sample the model (Fig. 1). ▷ Sampled from the Simulator.
4: Calculate summary statistics.
5: Calculate Euclidean distances between simulated and observed summary statistics.
6: Accept M data sets with the smallest Euclidean distances.
7: Return M accepted data sets.

Following the rejection algorithm, we applied the linear regression correction to compare ABC performance with and without linear regression correction. The input of the ABC-linear regression is the output of the ABC-rejection. In ABC-rejection we assign weights w_m to accepted data sets from ABC-rejection [41]. The weight w_m for each of accepted pairs from Algorithm 1 output are calculated using kernel $\kappa_{d_M}(d_m)$ [42], where d_m is the mth smallest Euclidean distance between standardized accepted summary statistics ($\boldsymbol{SumStat^*_{m\ scaled}}$) and standardized observed summary statistics ($\boldsymbol{SumStat_{obs\ scaled}}$), and $d_M > 0$ is the bandwidth parameter [43], in this case the largest of Mth Euclidean distance, order$\langle d_1, d_2, \ldots, d_{nsim}\rangle[M]$. The purpose of the kernel weights calculations is to apply in the calculation of weighted least squares regression coefficients $\hat{\boldsymbol{\beta}}_{WLS}$ for the linear regression correction analysis in Algorithm 2, where the accepted data set with smallest Euclidean distance d_i from Algorithm 1 output is adjusted the least and the accepted data set with largest d_i out of the M accepted is adjusted the most. The beta estimates vector is an approximate draw from the posterior. In our ABC-linear regression correction, the user can choose to calculate the kernel weights either based on Gaussian [41,44], or Epanechnikov [45] kernel and we later show in the results that the two kernels are similarly as effective. Further details on kernels, kernel weights, weighted least squares coefficients, and adjustment of accepted data sets parameter estimates are described in Appendix A.

The ABC-linear regression correction steps are shown in Algorithm 2.

Algorithm 2 ABC-linear regression correction algorithm performed on the output of the ABC-rejection. Input: standardized observed summary statistics, M accepted standardized summary statistics with corresponding M accepted parameters from ABC-rejection.

1: Input: standardized observed summary statistics, M accepted standardized summary statistics with corresponding M accepted parameters.
2: Calculate kernel weights. ▷ Calculate each of M weights based on kernel type.
3: Calculate weighted least squares regression coefficients.
4: Return adjusted parameter values from the linear regression adjustment.

5. Results

5.1. Overview

Below we describe two methods for estimating strength of selection under four migration (m) and mode of reproduction (sex) combinations, and 0,1, or 2 loci under selection combinations.

In the 1stmethod we describe, t^* and t_{final} are fixed 5 and 19 respectively, and the selection coefficients conditional on locus span on $s_i|I = i \sim$ Unif[-0.25,0.25] (see Section 3.2). The parameter space for loci under selection is at genomic locus corresponding to SNP number that can only take on values on $\{L/3, L/2, 2L/3\}$. An advantage for this method is no need for the initial summary statistic outlier scan (as described in Section 4.1) which would require simulation of $nsim = 100,000$ data sets per observed data set, with parameter space of potential loci under selection reduced to those identified by the outlier scan. In Section 5.2 and Section 5.2.1 we describe and perform an ABC analysis for each of total of $n_{ABC} = 10{,}000$ observed data sets, where we re-use simulated same $nsim = 100,000$ data sets for each analysis. We plot observed $signF_{ST}$ summary statistics for all 10,000 observed data sets and visually represent how the parameter space for the strength of selection differs depending on migration and mode of reproduction combinations (Fig. 3), and depending on number of loci under selection (Fig. 4). To reduce the complexity of variable parameters, we visually verify of this, simplified method case that ABC-linear regression correction estimation outperforms ABC-rejection estimation, and that there is no difference in performance of ABC-linear regression with Gaussian kernel versus with Epanechnikov kernel (**Appendix C**).

In the 2nd method, we performed four scenarios total of t^*, t_{final}, number of SNPs L, recombination rate r, and parameter space of selection conditional on locus $s_i|I = i$ combinations (see Table 2), unlike where in 1st method we performed only one combination. Additionally, we build on the technique from the 1stmethod by allowing the loci under selection of the observed data to span anywhere between locus corresponding to SNP 1 to L, then performing outlier scan as described in Section 4.1 which requires simulation of $nsim = 100,000$ data sets per observed data set. Based on the results from the 1st method about ABC-rejection versus ABC-linear regression, and corresponding ABC-linear regression two kernels, we performed analysis only with ABC-linear regression (after the ABC-rejection but without comparing to ABC-rejection) and compare ABC performance with four summary statistics combinations instead of one like in the 1st method, and found that $signF_{ST}$ summary statistics in ABC analysis is has low squared-bias and variance between observed and median accepted posterior selection parameters relative to the other three summary statistics combinations (Fig. 9).

5.2. Observed data, model simulations, and ABC for fixed potential loci under selection

In this 1st method, we assessed how the signal of selection from summary statistics changes when the mode of reproduction, strength of migration, as well as number of loci under selections change. As Wright's fixation index [25] measures the allele frequency differentiation among sampled populations [26] of sequenced data, and demographic history of *Saccharomyces cerevisiae* has been reported to play a role in gene expansion and contraction based on phylogeny reconstruction [46], here we investigate this demographic history and its effect on signature of selection.

For this, we assessed signal strength equivalent to scenario 1 for the 2nd method from Table 2 but for 10,000 ABC iterations ($n_{ABC} = 10,000$) and without the initial outlier scan on summarized observed data. This method has a computational time advantage as instead of simulating unique 100,000 ($nsim = 100,000$) data sets per one observed data, the simulated data sets are re-used. We present the results with $signF_{ST}$ summary statistics based on ABC-linear regression with Gaussian kernel (see Fig. 5, and Fig. 6), but we also compare performance of Gaussian versus Epanechnikov kernels, Gaussian kernel versus rejection, and Epanechnikov kernel versus rejection (see **Appendix C**).

The details of the experimental design of parameter space are described below.

5.2.1. Experimental design of parameter space

The experimental design of the parameter space which was the same as for scenario 1 in Table 2 is as follows: for $L = 100$ SNPs, with recombination rate of 3.0×10^{-4} Morgans/bp. The expected number of crossovers was $\mu_{n_r} = 4.95$ and average spacing of polymorphic sites on the genome every $SNP_{spacing} = 165$ sites as those we used in scenario 4 (Table 1, and Table 3 in Appendix A), resembling a biological yeast system where YPS128 and DBVPG1106 yeast strains of 12Mb genome length differ by over 70,000 SNPs or 1 SNP per $\sim$ 165 bp [8], as described in Section 2.1. The migration rate was randomly chosen to be 0.0, 0.2, or 0.5, sexual or asexual reproduction during migration generation. The migration generation cycles ($n_{cycles} = 4$) took place every fifth generation ($t^* = 5$) for total of $t_{final} = n_{cycles}t^* - 1 = 19$ generations. Both, the observed and simulated data can have either 0, 1, or 2 loci under selection with prior parameter space of selection coefficients conditional on locus $s_i|I = i \sim$ Unif$[-0.25, 0.25]$.

The experimental design part that differed from scenario 1 was: We performed 10,000 ABC iterations instead of 100. Out of the potential 0, 1, or 2 loci under selection, the possible loci that the selection could act on were $\{L/3, L/2, 2L/3\}$, more specifically $\{33^{\text{rd}}$ SNP, 50^{th} SNP, 67^{th} SNP$\}$. For this small set of the only possible loci under selection, no outlier scan on observed summarized data was performed, which enabled us to re-use the same 100,000 simulated data sets for the ABC iterations.

The experimental design of the parameters space for selection coefficients and rates of migration, with possible models of reproduction during migration generations of the model are seen in Fig. 2.

For the signature of the strength of selection due to mode of reproduction and strength of migration, we have evaluated $signF_{ST}$ summary statistic of each of 10,000 observed data sets under four variable modes of reproduction (sex) and strength of migration

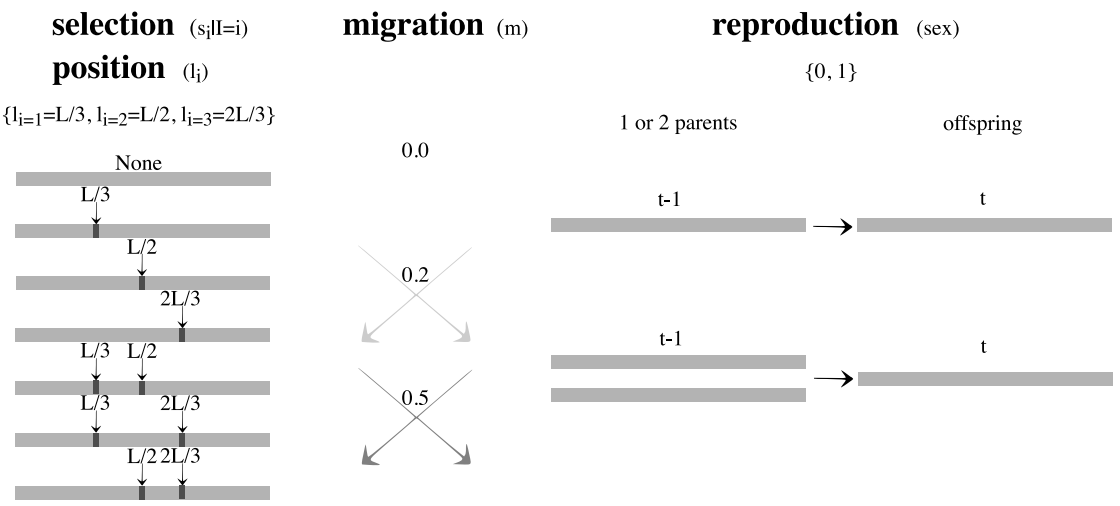

Fig. 2. Possible parameter space of selection coefficients for the 1st method, i.e. without the initial outlier scan on summarized observed data. Scheme of selection coefficients conditional on genome loci ($s_i|I = i$), rates of migration (m), and possible modes of reproduction during migration generations (sex). The positions under selection can only take on values on $\{L/3, L/2, 2L/3\}$, being able to re-use the same $nsim = 100,000$ data sets for the ABC iterations.

(m) combinations: $sex = 0$ and $m = 0$, $sex = 1$ and $m = 0$, $sex = 1$ and $m = 0.2$, $sex = 1$ and $m = 0.5$, while equally probable seven combinations of 0,1,2 loci under selection within a set of $\{L/3, L/2, 2L/3\}$ possible SNP loci (see Fig. 2). In Fig. 3 we see a visible pattern in an increase in summary statistics values away from 0 ($signF_{ST} = 0$ when the fixation index $F_{ST} = 0$) at selected loci with a decrease in migration rate, and increase in genetic hitchhiking effect with an asexual mode of reproduction ($sex = 0$) given same m of $m = 0.0$ as the recombinational distance from the loci under selection is absent.

For the signature of strength of selection due to a variable number of loci under selection, we have evaluated $signF_{ST}$ summary statistic of each of 10,000 observed data sets where 0, 1, and 2 loci are under selection within a set of $\{L/3, L/2, 2L/3\}$ possible SNP loci, while equally probable four combinations of mode of reproduction and migration described above (see Fig. 2). In Fig. 4 we see a visible pattern in summary statistics values closest to 0 for no loci under selection, increase in the magnitude of summary statistics values away from 0 at single locus under selection, and in-between the magnitude of summary statistics values away from 0 and increase in genetic hitchhiking effect for two loci under selection.

In Fig. 3 we see a break down of LD when recombination rate is present versus absent, given no migration rate, and in Fig. 4 we see more break down in LD for one locus under selection instead of two loci, given same average genomic recombination rate, which is consistent with recombination as the primary source of LD break down [7].

To further examine the signature of LD decay and distance between loci under selection, given fixed average genomic recombination rate, we looked at cases with loci under selection in closer proximity to each other. We evaluated effect of number of loci and distance between loci under selection for two loci cases, i.e. four combinations: none, one locus, two loci $L/6$ distance apart, and two loci $L/3$ distance apart, as well as each of seven loci combinations, i.e: none, $L/3$, $L/2$, $2L/3$, $(L/3, L/2)$, $(L/3, 2L/3)$, $(L/2, 2L/3)$ but we could not distinguish visually further differences than those seen in Fig. 4 in observed summary statistics values due to position (seven combinations), nor the distance (four combinations) (see **Appendix C**).

In order to answer the question how well the signal we recover the signal of strength of selection, with variable selection coefficients priors, when the mode of reproduction, strength of migration, and number of loci under selection change, we evaluated the bias and the variability in the estimates of selection coefficients from the ABC. For this, we plotted the true observed selection coefficients versus: posterior medians, MSEs between true observed versus median of posteriors, variance of the posterior medians, and bias squared between true observed and posterior medians based on the ABC-linear regression Gaussian kernel. We picked Gaussian kernel because the PoPoolation software – a pipeline for analyzing pooled next generation sequencing data [47] – uses Gaussian kernel smoothing [48]. We assessed them based on four modes of reproduction and migration combinations, and based on number of loci under selection. For the mode of reproduction and migration (Fig. 5), we see a positive relationship between the magnitude of true observed selection coefficient values for strongest migration rate ($m = 0.5$) and variance as well bias (squared), with high migration rate contributing to opposing the effect of divergent selection. For the same range of selection coefficient values, the range of summary statistic values is small of data generated with high m, in compare to low m (Fig. 3). The signal of the strength of divergent selection also diminishes with number of loci under selection. In (Fig. 6), we see a positive relationship between the magnitude of true observed selection coefficient values for two loci, and variance as well bias (squared).

We also assessed Gaussian kernel performance with respect to Epanechnikov, ABC-rejection (Algorithm 1) versus Gaussian, and ABC-rejection (Algorithm 1) versus Epanechnikov. We compared to Epanechnikov kernel, and ABC-rejection and found no visual difference between kernel types, but an improvement in selection coefficient estimations for both Gaussian and Epanechnikov in comparison to ABC-rejection (**Appendix C**).

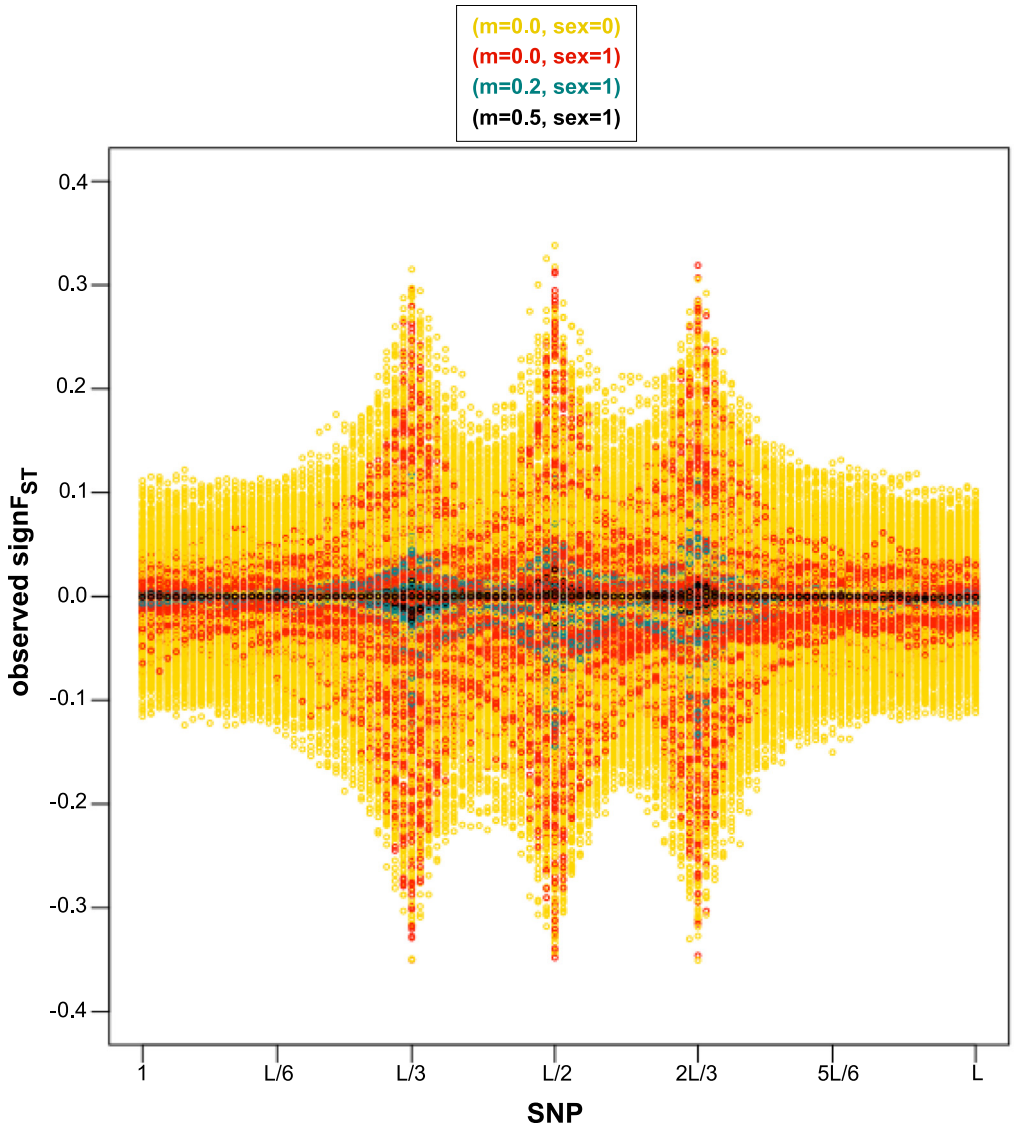

Fig. 3. SNP loci vs. $signF_{ST}$ summary statistics for the 1st method, i.e. without the initial outlier scan. Plots for each of 10,000 simulator output data sets assumed as observed under four of migration (m) and mode of reproduction (sex) combinations selected randomly with equal probability: $sex = 0$ and $m = 0.0$ in yellow, $sex = 1$ and $m = 0.0$ in red, $sex = 1$ and $m = 0.2$ in turquoise, $sex = 1$ and $m = 0.5$ in black, and selected randomly with equal probability scenarios of loci under selection as seen in Fig. 2. A visible pattern is shown of an increase in summary statistics values away from 0 at selected loci with a decrease in migration rate, and an increase in genetic hitchhiking effect with an asexual mode of reproduction ($sex = 0$) (absence of recombination) given same m of $m = 0.0$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

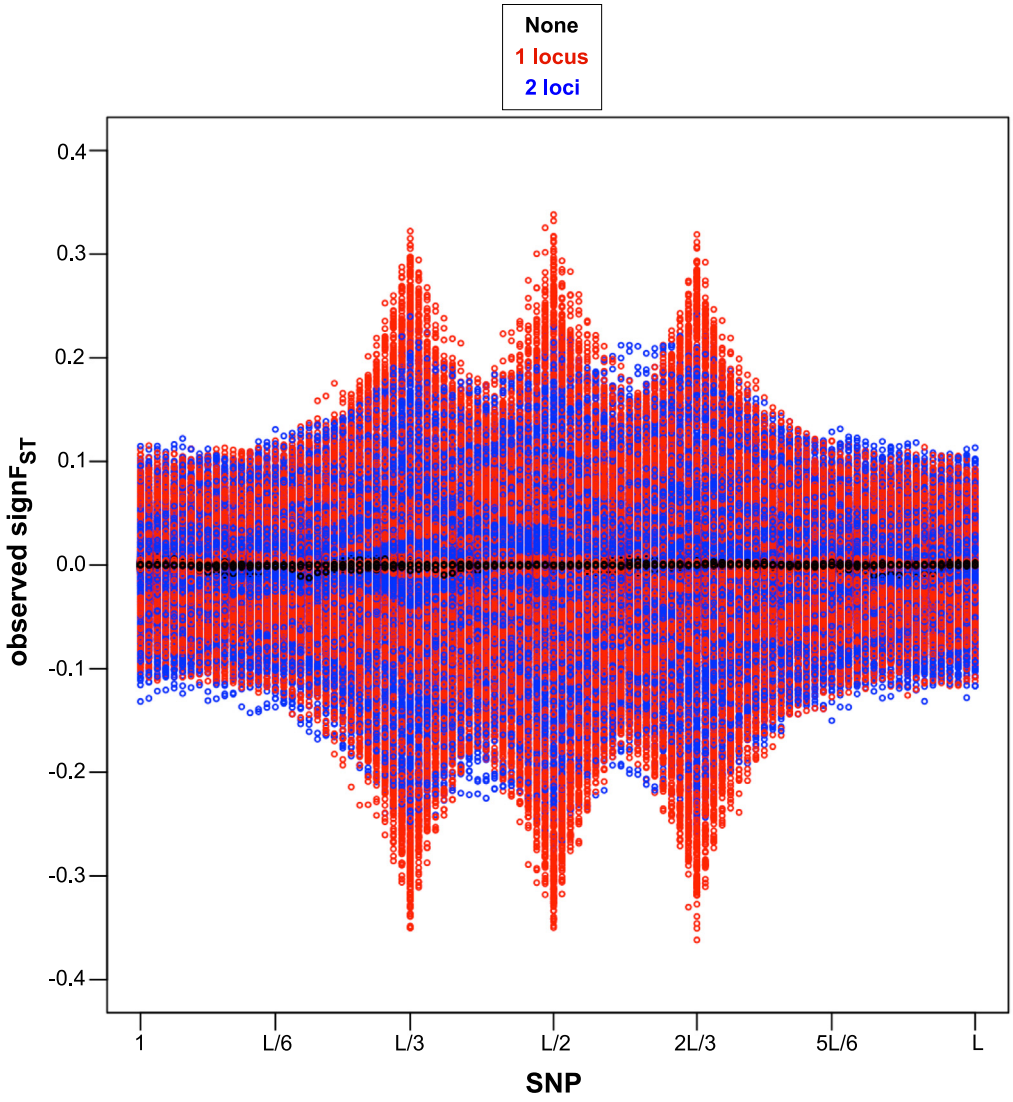

Fig. 4. SNP loci vs. $signF_{ST}$ summary statistics for the 1st method, i.e. without the initial outlier scan. Plots for each of 10,000 simulator output data sets assumed as observed under four of migration (m) and mode of reproduction (sex) combinations selected randomly with equal probability scenarios of loci under selection: none in black, one locus ($L/3$, or $L/2$, or $2L/3$) in red, two loci ($L/3$ with $L/2$, or $L/3$ with $2L/3$, or $L/2$ with $2L/3$) in blue, as seen in Fig. 2. A visible pattern is shown of summary statistics values closest to 0 for no loci under selection, largest magnitude in summary statistics values away from 0 at single locus under selection, and in-between the magnitude of summary statistics values away from 0 and increase in genetic hitchhiking effect for two loci under selection (lower observed frequency of recombination events between two loci under selection at a confined distance apart). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

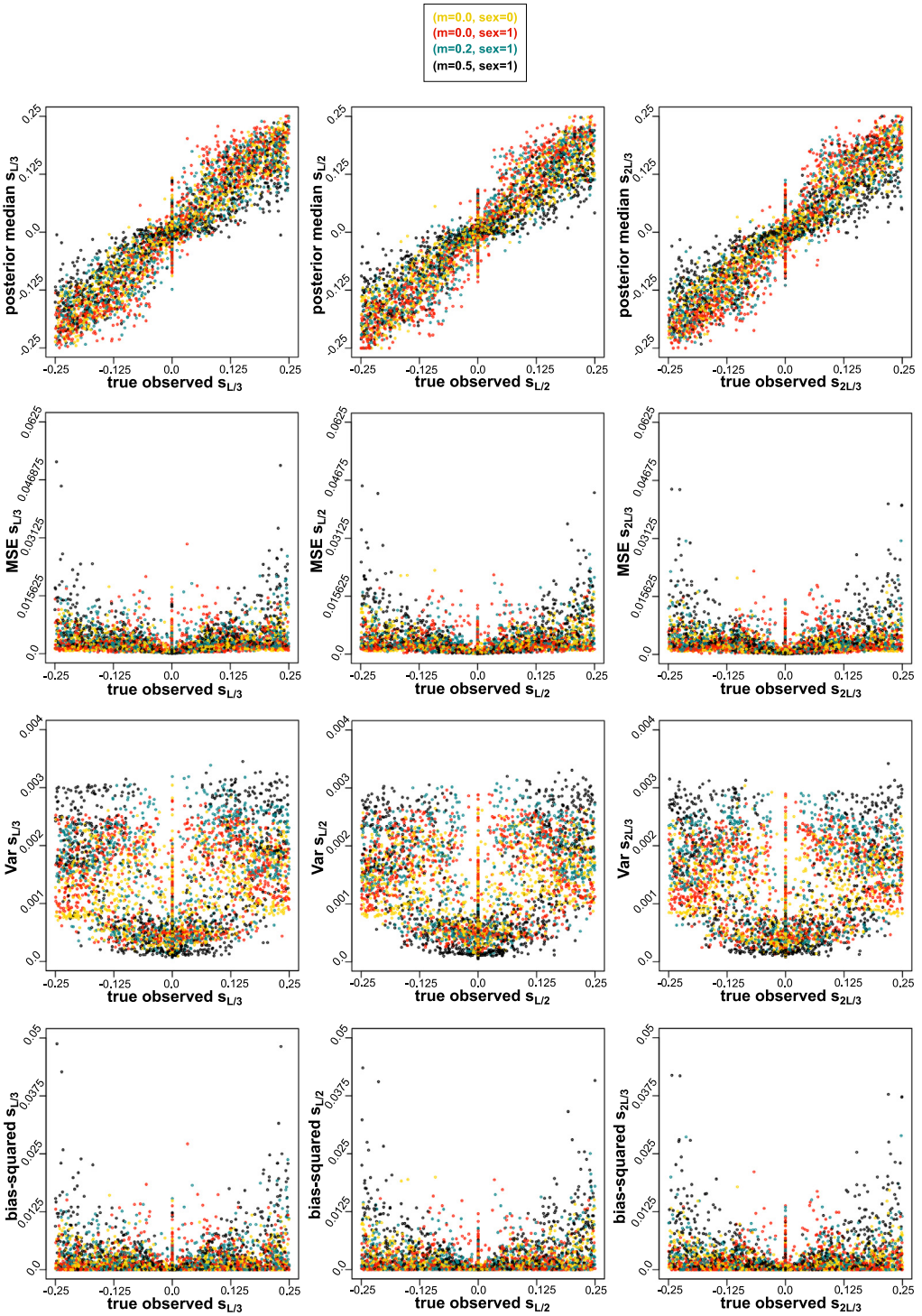

Fig. 5. True parameter value under which the observed parameter is generated (x-axis) vs. median, MSE, variance, bias-squared for $(L/3)$th, $(L/2)$th and $(2L/3)$th SNP respectively for the 1st method, i.e. without the initial outlier scan, from $n_{ABC} = 10{,}000$ ABC tests, based on $signF_{ST}$ summary statistics from Algorithm 2 with Gaussian kernel. Colors based on four migration (m) and mode of reproduction (sex) combinations of observed data sets: $sex = 0$ and $m = 0.0$ in yellow; $sex = 1$ and: $m = 0.0$ in red, $m = 0.2$ in turquoise, $m = 0.5$ in black. A pattern is shown of a positive relationship between the magnitude of true observed selection coefficient values for highest migration rate $(m = 0.5)$, and variance as well bias (squared), with decreased signature of selection at high migration rate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

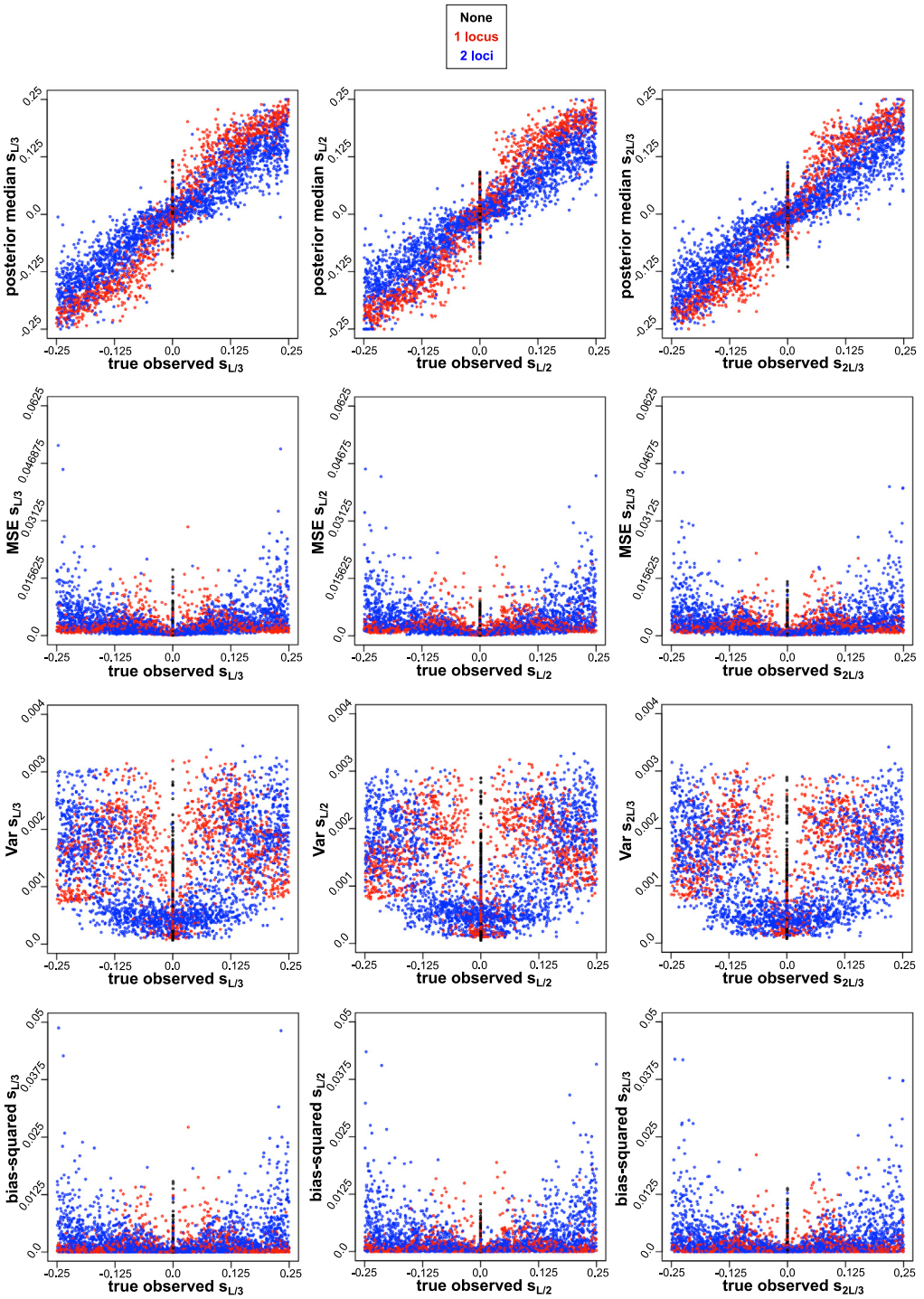

Fig. 6. True parameter value under which the observed parameter is generated (x-axis) vs. median, MSE, variance, bias-squared for $(L/3)$th, $(L/2)$th and $(2L/3)$th SNP respectively for the 1st method, i.e. without the initial outlier scan, from $n_{ABC} = 10{,}000$ ABC tests from $signF_{ST}$ summary statistics from Algorithm 2 with Gaussian kernel. Colors based on three combinations of number of loci under selection of observed data sets: none in black, one locus in red, and two loci in blue. A pattern is shown of a positive relationship between the magnitude of true observed selection coefficient values for two loci, and variance as well bias (squared), with greater genetic hitchhiking effect on two loci under selection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.3. Observed data, model simulations, and ABC for potential loci under selection identified by outlier scan

In the 2nd method, to answer question how expansion of the parameter space for the positions of loci under selection contributes to accuracy in estimation of selection coefficients, we performed ABC evaluations for estimating strength of selection for model scenarios described in Table 2. The observed data sets had randomly selected 0, 1, or 2 loci under selection, and randomly selected loci under selection along SNP loci 1 to L (see **Model** section on parameter methodology). We identified candidates for loci under selection on the n_{ABC} observed data sets via the outlier scan using the F_{ST} summary statistic (see Section 4.1) first, followed by the model simulations, then ABC.

We performed a total of four scenarios. For scenarios 1–3, we simulated same number of SNPs, average recombination rate, as described in 1st method in sub- Section 5.2.1, and with same expected number of crossovers and average spacing of polymorphic sites on the genome equivalent to the biological yeast system as in scenario 4 (Table 1, and Table 3 in Appendix A). Because of relatively low number of SNPs simulated, which referred which required lower computational time per simulated data set, we performed 100 ABC iterations (n_{ABC}), which translated into 100 observed data sets (one ABC iteration per one observed data set), with $nsim = 100,000$ unique data sets simulated conditional on outlier scan on potential locus under selection per each observed data set.

For scenario 1, the migration-selection cycle occurred every $t^* = 5$ generations, with total number of cycles $n_{cycles} = 4$, and final generation $t_{final} = n_{cycles}t^* - 1 = 19$. The prior parameter space of selection conditional on locus is $s_i|I = i \sim \text{Unif}[-0.25, 0.25]$. This parameter space implies that if an overall population fitness ω is $\omega = 1$ at $t = 0$, that is the average fitness of all carriers $\frac{1}{N_e}\sum_{n=1}^{N_e}\omega_n = 1$ at $t = 0$, for extreme strength of selection if a carrier of one locus under selection has $s_i|I = i = -0.25$ or $s_i|I = i = 0.25$, it has 25% lower or higher fitness respectively than the population average at $t = 0$. For two loci under selection, this translates to 50% lower or higher fitness respectively than the population average at $t = 0$.

For scenario 2, the only parameters that differ from scenario 1 is the length of the selection cycle $t^* = 50$, a 10-fold increase from scenario 1, and the prior parameter space of selection conditional on locus is $s_i|I = i \sim \text{Unif}[-0.025, 0.025]$, a 10-fold decrease from scenario 1, as one would expect fixation to reach slower with weaker selection.

For scenario 3, the only parameter that differs from scenario 2 is number of cycles $n_{cycles} = 2$, thus final generation $t_{final} = n_{cycles}t^* - 1 = 99$. This scenario is to test whether the tested number of mixing (migration) times plays a significant role in the signature of selection in comparison to scenario 2.

For scenario 4, the recombination rate resembles closer to a genomic recombination rate of sequenced yeast data. The average recombination rate of *Saccharomyces cerevisiae* has been reported 3.5×10^{-6} Morgans/bp in the literature [9], with inferred genome-wide recombination profiles from sequenced isolates from an advanced intercross line (AIL) to be as high as 3.0×10^{-5} Morgans/bp for a two-way cross at genome hotspots [10]. Here, for the scenario resembling sequenced yeast, we fixed genome recombination rate of $r = 2.0\times10^{-5}$ Morgans/bp, a realistic value for recombination hotspots in yeast [10]. The simulated number of SNPs for scenario 4 for the genomic chunk is $L = 1,500$, around an average number of SNPs of a yeast cross for one-third a chromosome [8], which still holds true that the expected number of cross-over events is $\mu_{n_r} = 4.95$ (Table 1, and Table 3 in Appendix A) as for scenarios 1–3, and as for the non-outlier scan earlier scenario with smaller parameter space of potential loci under selection ($\{L/3, L/2, 2L/3\}$). The pior parameter space of length of selection cycles, number of cycles (and thus final generation), and selection coefficients conditional on loci are equivalent to scenario 2 ($t^* = 50$, $n_{cycles} = 4$, $t_{final} = n_{cycles}t^* - 1 = 199$, $s_i|I = i \sim \text{Unif}[-0.025, 0.025]$). Due to longer computational time because we simulated more SNPs in scenario 4 than in scenarios 1–3, we picked 10 instead of 100 data sets as observed and thus performed $n_{ABC} = 10$ instead of $n_{ABC} = 100$ ABC iterations.

Full list of scenarios with initial outlier scan for which the evaluations were performed are described in Table 2. The Fig. 7 and Fig. 8 show the summary statistics of observed data sets based on mode of reproduction and migration, and number of loci under selection respectively for each of the four scenarios. The ABC performance of MSE, variance and squared-bias between true selection parameter values under which the data are generated, and the values of posterior distributions for each of the four scenarios are shown in Fig. 9. In ABC analysis estimators are set to medians of the posterior distributions.

Table 2
Scenarios under which the initial parameters varied for the model simulations and ABC, with parameter values in blue that are unique to particular scenario, and those in black that are shared with at least one other scenario.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4 (sequenced yeast data)
Varying Parameters	$L = 100$, $r = 3.0\times10^{-4}$, $t^* = 5$, $t_{final} = 19$, $s_i\|I = i \sim$ Unif$[-0.25, 0.25]$	$L = 100$, $r = 3.0\times10^{-4}$, $t^* = 50$, $t_{final} = 199$, $s_i\|I = i \sim$ Unif$[-0.025, 0.025]$	$L = 100$, $r = 3.0\times10^{-4}$, $t^* = 50$, $t_{final} = 99$, $s_i\|I = i \sim$ Unif$[-0.025, 0.025]$	$L = 1,500$, $r = 2.0\times10^{-5}$, $t^* = 50$, $t_{final} = 199$, $s_i\|I = i \sim$ Unif$[-0.025, 0.025]$
Total number of ABC iterations (n_{ABC})	100	100	100	10
Total number of simulated data sets for all ABC iterations ($n_{ABC} \times nsim$)	10 million	10 million	10 million	1 million

Out of the four scenarios, we see strongest pattern of observed summary statistics for scenario 1, where the range of selection coefficients is 10-fold greater, with 10-fold shorter migration-selection cycles, characterized by sharp peaks (Fig. 7 and Fig. 8). We also see that the observed $XP-EHH$ is almost exclusively positive. From expression (8) with population $\boldsymbol{X}$ in the numerator and

population $\boldsymbol{Y}$ in the denominator, population $\boldsymbol{X}$ and $\boldsymbol{Y}$ are modeled under selection and under neutrality respectively. The numerator is larger due larger sum of the extended haplotype homozygosity decay around the locus the selection is acting on. LD-based statistic $XP{-}EHH$ has been shown to be more effective when one variant of allele is near fixation within one population [49,50]. In scenario 1 with stronger selection, and shorter migration-selection cycles, the genetic drift is expected to have smaller effect proportionally to strength of selection and length of migration-selection cycles. With absent migration ($m = 0$), a build-up of selection over time is expected to occur, and a break down of LD when recombination is present ($sex = 1$) instead of absent ($sex = 0$), shown most clearly in our results with $XP - EHH$.

For scenario 4 (resembling a biological yeast system), where 15-fold larger number of SNPs are simulated than for scenarios 1–3, we see peaks of observed summary statistics values with more stochastic effect between the SNPs unlike a more smooth pattern in observed summary statistics of neighboring SNPs in scenarios 1–3. This pattern resembles more so the pattern of the level of heterozygosity [51] on domestication and divergence of *S. cerevisiae*, and less so in scenarios 1–3.

Our results presented show some difference in genetic divergence dependent on number of mixing cycles. We see larger range of observed summary statistics values for scenario 2 compared to scenario 3, with four migration-selection cycles instead of two. The observed summary statistics values deviate the most from zero for data generated under no migration in both scenarios, where the process of genetic drift and/or divergent selection affects the accumulation genetic differences [52].

Besides the differences in patterns of observed summary statistics between the scenarios, we see a consistent pattern across the scenarios. The lowest level of divergence, expressed as the lowest magnitude of observed summary statistics, is seen for the strongest rate of migration for all summary statistics (Fig. 7), and no loci under selection (Fig. 8) for $signF_{ST}$ and $(p_X - p_Y)$.

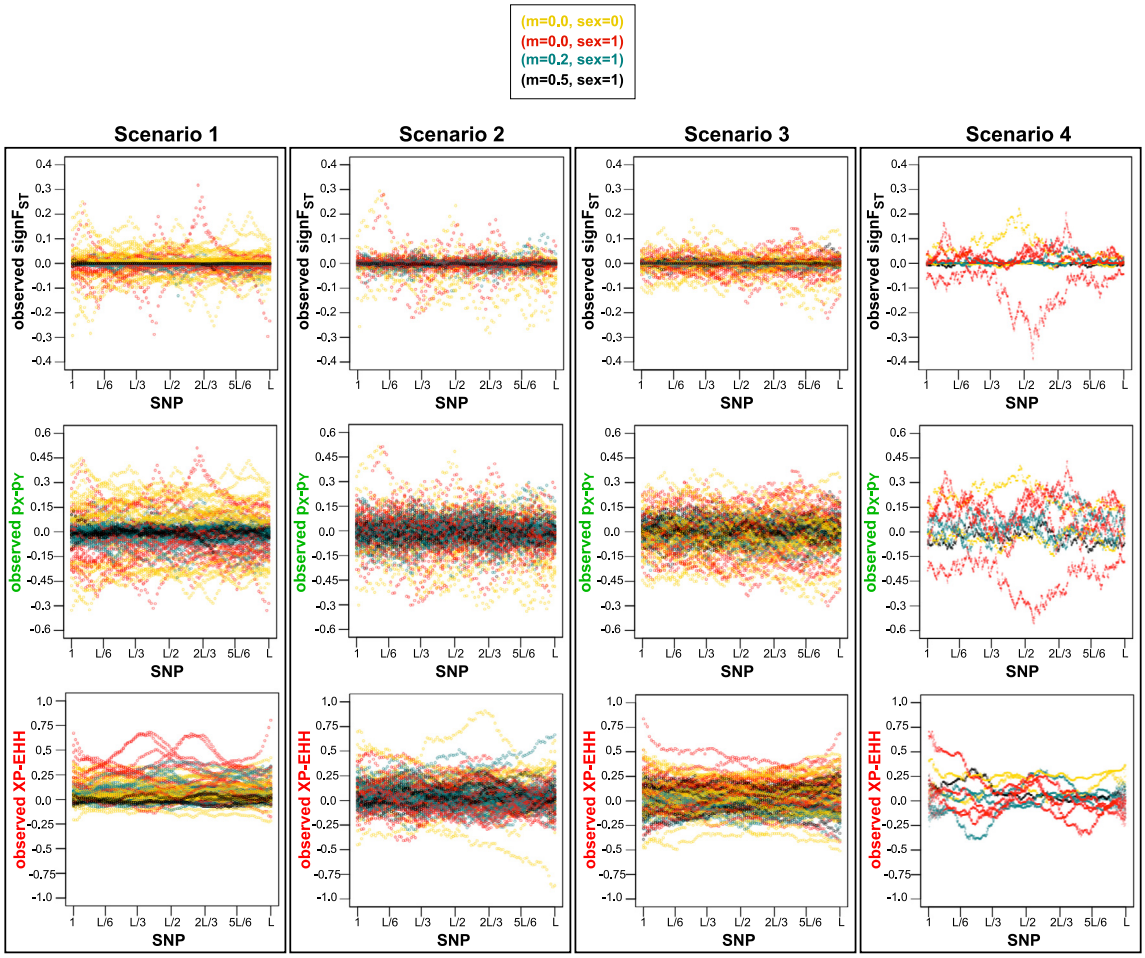

Fig. 7. SNP loci vs. observed summary statistics for the 2nd method, i.e. each of four scenarios with initial outlier scan, (Table 2). Simulator output data sets assumed as observed under four of migration (m) and mode of reproduction (sex) combinations selected randomly with equal probability: $sex = 0$ and $m = 0.0$ in yellow; $sex = 1$ and: $m = 0.0$ in red, $m = 0.2$ in turquoise, and $m = 0.5$ in black. Data selected randomly with equal probability scenarios of loci under selection. A visible pattern is shown of summary statistics values closest to 0 for the strongest migration rate of $m = 0.5$ across all scenarios, constraining the genetic divergence. Clear summary statistics peaks shown in scenario 1, with smaller effect of genetic drift on strong selection, and with $m = 0$ most clear pattern of LD break down when $sex = 1$ instead of $sex = 0$ seen with $XP - EHH$. Pattern of more stochastic effect between the peaks of summary statistics values between neighboring SNPs seen in scenario 4, resembling more of heterozygosity pattern seen on a *S. cerevisiae* genome. A larger range of summary statistic values seen for scenario 2 compared to scenario 3, with four migration-selection cycles instead of two. Colors of observed summary statistics labels correspond to colors of plots in Fig. 9. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

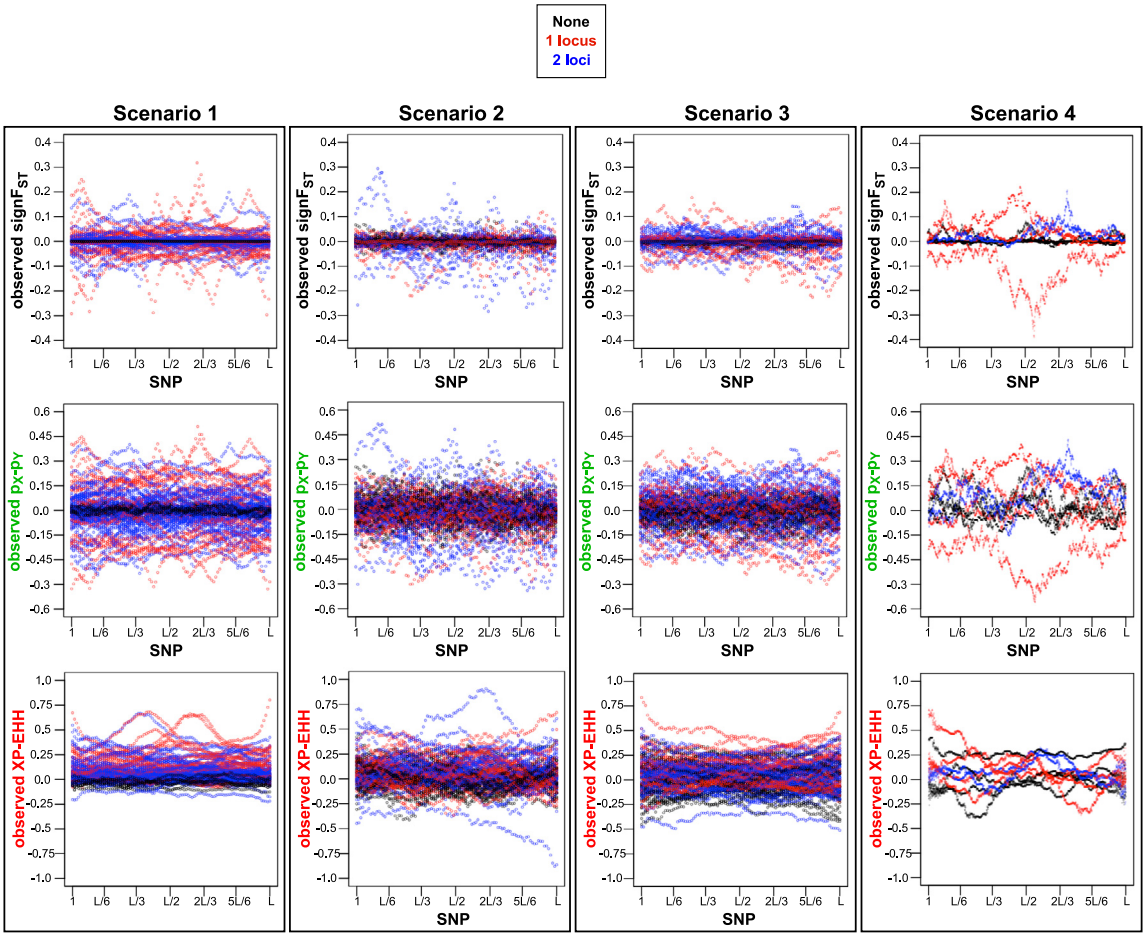

Fig. 8. SNP loci vs. summary statistics for the 2nd method, i.e. each of four scenarios with initial outlier scan, (Table 2). Simulator output data sets assumed as observed under four of migration (m) and mode of reproduction (sex) combinations selected randomly with equal probability scenarios of number loci under selection occuring within the loci identified by the outlier scan: none in black, one in red, and two blue. Clear summary statistics peaks shown in scenario 1, with smaller effect of genetic drift on strong selection. Pattern of more stochastic effect between the peaks of summary statistics values between neighboring SNPs seen in scenario 4, resembling more of heterozygosity pattern seen on a *S. cerevisiae* genome. A larger range of summary statistic values seen for scenario 2 compared to scenario 3, with four migration-selection cycles instead of two. A visible pattern of $signF_{ST}$ and $(p_X - p_Y)$ values closest to 0 for no loci under selection across all scenarios. Colors of observed summary statistics labels correspond to colors of plots in Fig. 9. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

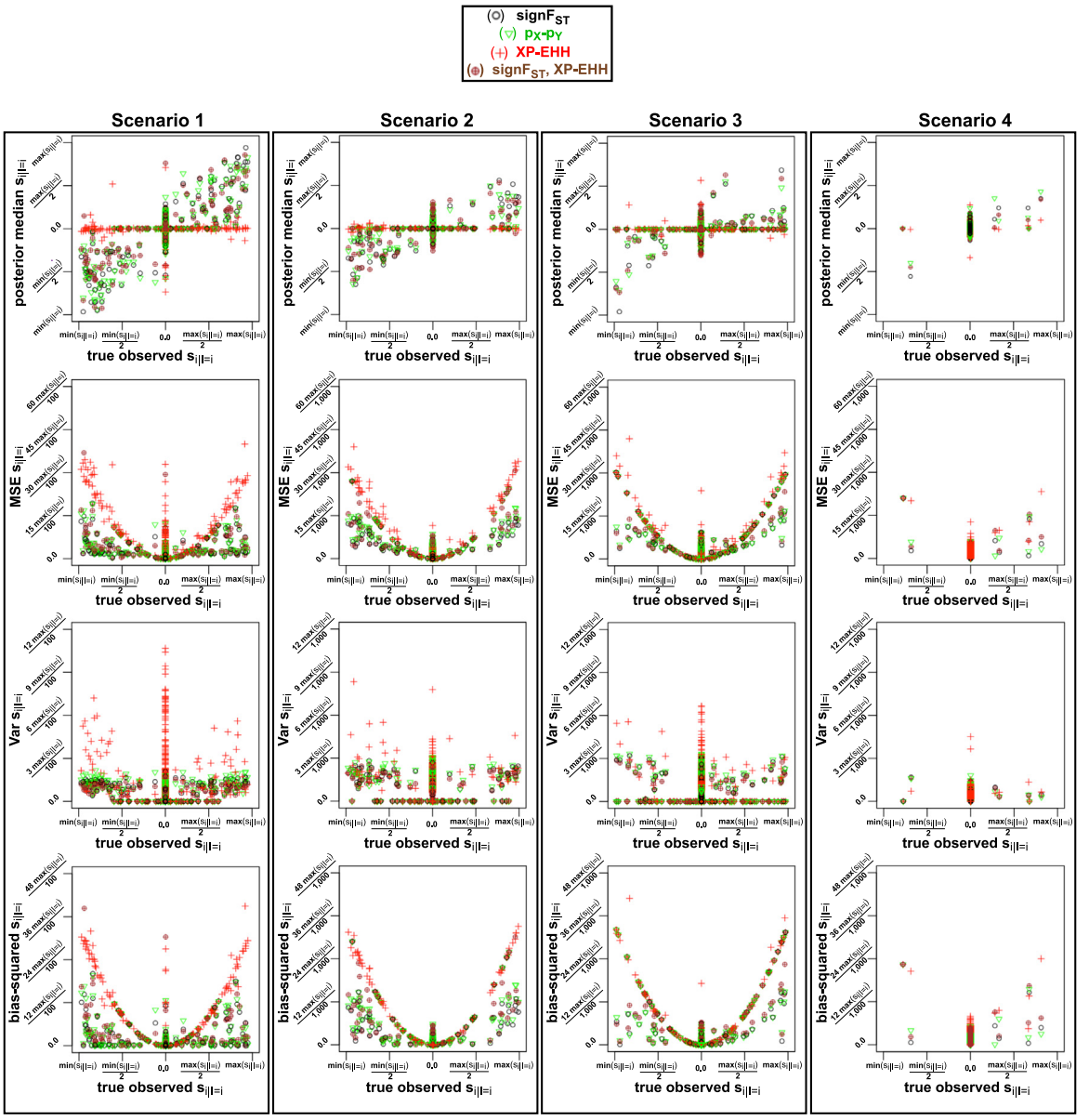

Fig. 9. True parameter value under which the observed parameter is generated (x-axis) vs. median, MSE, variance, bias-squared for loci corresponding to SNP for the 2nd method, i.e. determined by initial outlier scan, followed by simulation of data sets and ABC from Algorithm 2 with Gaussian kernel correction. Colors based on four combinations of summary statistics inputted to ABC are described in the figure legend. Plots for four scenarios, where in scenario 1 the range of selection coefficient is 10-fold larger ($[-0.25, 0.25]$): thus the y-axis of the posterior median is 10-fold larger, the MSE, variance and bias-squared is (10-fold)-squared larger than for scenarios 2–4. A pattern is shown of posterior median parameter values closest to true observed values based on $signF_{ST}$, $(p_X - p_Y)$, and $(signF_{ST}, p_X - p_Y)$ for scenario 1, and least information captured with $XP-EHH$.

After examining patterns of observed summary statistics by mode of reproduction with migration, and by number of loci under selection, we compared the ABC performance with these observed summary statistics by evaluating the variance of ABC posteriors, squared-bias between of the posterior estimates and true observed parameters. We find a clear pattern of posterior estimates to be less biased for scenario 1, where the range of selection coefficients is 10-fold of those in scenarios 2–4, and where the migration-selection cycles are shorter (Fig. 9). Once explanation could be that short migration-selection cycles do not allow for the significant build-up of genetic drift, and/or that strong selection has too significant of effect for the genetic drift to act on the population. We also determine a pattern of $XP-EHH$ capturing least information about the signal of divergent selection, and discuss potential explanation in Section 6.2.

6. Discussion

6.1. Application

By combining ABC methods that incorporate summary statistics on large simulation study from developed simulator, we present an approach of estimating model parameters, which can bypass evaluation of exact likelihood function. We show it on simulations with variable migration rates, modes of reproduction, and number of loci under selection, where fixation index summary statistics outperformed cross-population extended haplotype homozygosity in terms of precision and accuracy. We also recommend fixation index over the LD-based statistic cross-population extended haplotype homozygosity, because calculating the linkage disequilibria is computationally expensive [53].

Moreover, when considering evolutionary models of complex systems, they provide valuable insights into how systems adapt, change, and evolve [54]. These models, by simulating intricate interactions among diverse components, unlock crucial patterns shaping the adaptive nature of systems, thereby laying the groundwork for understanding where the strength of recovered signals lies in representing divergent selection. An important question about the application of these methods is: for which part of the parameter space of the model the recovered signal about the selection coefficients works well? In Fig. 9 scenario 1 for instance the posterior median estimates of selection coefficients are closer to the values of the true observed selection coefficients in compare to scenarios 2–4. In scenario 1, the migration-selection cycles are 10-fold smaller, and the selection coefficients range is 10-fold larger thus affected less by the genetic drift, supporting that selection can only be assessed if it is high enough to outperform the effect of genetic drift [55]. The shorter migration-selection cycles with overall less generations simulated, took less time to simulate. We showed in scenario 4 that our model is scalable to recombination rate parameter resembling sequenced yeast data [10], with same expected number of recombination events Table 1, Table 3 in Appendix A) as for scenarios 1–3. If the parameter space of possible loci under selection can be assumed to those in Fig. 2, a more robust number of ABC iterations on data resembling biological yeast recombination rate is feasible in terms of computational time.

6.2. Recombination rate

A limitation of the study where the main focus was estimation of selection coefficients, was fixing the recombination rate to an expected number of 4.95 recombination events (Table 1, Table 3 in Appendix A) for recombination rates 3.0×10^{-4} Morgans/bp and resembling biological data 2.0×10^{-5} Morgans/bp [10] for 100 and 1500 SNPs respectively. We fixed the recombination rate to same for all simulations, such that the parameter r represents the average recombination rate on the simulated genome section. We explain out motive for fixing the recombination rate below.

With our model we attempted to estimate strength of selection with variable recombination rates of simulated data sets, however, we were unable to get consistent estimates. While the buildup of LD (i.e. the correlation between nearby variants of alleles as opposed to random association of alleles [56]) can be a result of several population genetic forces, recombination is the only primary method to break it down [7]. The absence of recombination between sites under selection will reduce the overall effectiveness of selection, a phenomenon known as the Hill-Robertson (HR) effect [7,57,58]. Our main focus was estimation of strength of selection, with successfully applied variability in mode of reproduction, and in strength of migration, but not with variability in recombination rate. As the $XP-EHH$ was least effective in estimation of selection with fixed recombination, and $XP-EHH$ has been used to measure the haplotype lengths between two populations [32,33], we would expect $XP-EHH$ to perform better in estimation of recombination without taking into account varying strength of selection, mode of reproduction and migration combined.

Future work direction would be an exploration of variable recombination rates.

CRediT authorship contribution statement

Martyna Lukaszewicz: Developed the theoretical model, Wrote the algorithm for the population genetics model, Applied the approximate Bayesian computational methods with testing different summary statistics, Explored the parameter space for which strong signal of model is recovered. **Ousseini Issaka Salia:** Performed the bioinformatics data cleaning and analysis to estimate the recombination rate for the sequenced yeast strains. **Paul A. Hohenlohe:** Presented the idea of estimation of divergent selection in biological system, for which received the grant funding, Encouraged Martyna Lukaszewicz after identifying optimal parameter space to recover divergent selection estimates with scaling the recombination rate to biologically relevant phenomena. **Erkan O. Buzbas:** Presented the idea of estimation of divergent selection in biological system, for which received the grant funding, Encouraged Martyna Lukaszewicz to perform computationally-efficient explanatory analysis for recovering divergent selection estimates under variable parameters.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Science Foundation (NSF) Grants DEB-1655809 and DBI-0939454, National Institutes of Health (NIH) COBRE Phase III Grant, United States P30GM103324, to PH and EB.

All authors discussed the results and contributed to the final manuscript.

Appendix A

see Tables 4 and 5

Algorithm 3 Algorithm for population process given in Fig. 1. Input: Effective population sizes N_e, for $\boldsymbol{X}$ and $\boldsymbol{Y}$; Genome size L, for each individual; Vector of loci of selected sites, $\boldsymbol{L_s}$; Vector of alleles at selected sites, a; recombination rate of parental genomes, r; strength of selection per locus s_i; migration rate from population $\boldsymbol{X}$ to population $\boldsymbol{Y}$, m_{XY}; migration rate from population Y to population $\boldsymbol{X}$, m_{YX}; migration event, between t^*-1 and t^*, SNP spacing on the genome, $SNP_{spacing}$ (assumed equally spaced SNPs). Parts I, II, A, B and C of Fig. 1 are performed on 2–5, 6–20, 21–30, 31–45, and 46–48 respectively.

Input: $N_e, L, \boldsymbol{L_s}, \boldsymbol{a^*} = (a_1^*, a_2^*, \cdots, a_{\|\boldsymbol{L_s}\|}^*), r, \boldsymbol{s^*} = (s_1^*, s_2^*, \cdots, s_{\|\boldsymbol{L_s}\|}^*), m_{XY}, m_{YX}, t_m, sex, SNP_{spacing}$
I
Build F_1:
Sample allele $a \sim$ Discrete Unif$[0,1]$ and replicate L times for each
genome carrier n, $n \in \{1, 2, \ldots, N_e\}$ in $\boldsymbol{X}_{\boldsymbol{F_1}}$ and $\boldsymbol{Y}_{\boldsymbol{F_1}}$ respectively
II
Build F_2:
for j in $\{\boldsymbol{X},\boldsymbol{Y}\}$ **do**
    for n in 1 to N_e **do**
        Sample with equal probability parent, $p_1 \sim$ Discrete Unif$[1, N_e]$ from $\boldsymbol{X}_{\boldsymbol{F_1}}$
        Sample with equal probability parent, $p_2 \sim$ Discrete Unif$[1, N_e]$ from $\boldsymbol{Y}_{\boldsymbol{F_1}}$
        Sample number of recombination events independently of each other,
        $n_r \sim$ Bin$(L \times SNP_{spacing} - 1, r)$
        Sample loci of recombination events,
        $\boldsymbol{L_r} \sim$ Discrete Unif$[n_r, L \times SNP_{spacing} - 1]$
        if $rv \sim$ Unif$[0,1] < 0.5$ **then**
            p_1 reassigned to p_2 and p_2 reassigned to p_1
        end if
    end for
end for
A
Selection, genetic drift, asexual reproduction at time t:
for j in $\{\boldsymbol{X},\boldsymbol{Y}\}$ **do**
    for n in 1 to N_e **do**
        Calculate absolute fitness for each individual: $\omega_n^{(j,t-1)} = \sum_{i=1}^{\|\boldsymbol{L_s}\|} \left(1 + \mathbf{I}_{n\{a_i \in (j,t-1)\}} s_i\right)$
    end for
    Sample offspring ${N_e}^{(j,t)} \sim$ Multinomial$(N_e, \boldsymbol{p})$,
    where $\boldsymbol{p} = \left(p_1, p_2, \ldots, p_N\right)$
    and $p_n = \frac{\omega_n^{(j,t-1)}}{\sum_{n=1}^{N_e} \omega_n^{(j,t-1)}}$
end for
B
Migration, $m_{XY} = m_{YX} = m$:
if $m > 0$ **then**
    Sample parents that migrate from $\boldsymbol{X}$ to $\boldsymbol{Y}$: $pID_{XY} \sim$ Discrete Unif$[1, N_e]$
    Sample parents that migrate from $\boldsymbol{Y}$ to $\boldsymbol{X}$: $pID_{YX} \sim$ Discrete Unif$[1, N_e]$
    Substitute leaving pID_{XY} to migrating pID_{YX} in $\boldsymbol{X}$
    Substitute leaving pID_{YX} to migrating pID_{XY} in $\boldsymbol{Y}$
else
    No migration
    for j in $\{\boldsymbol{X},\boldsymbol{Y}\}$ **do**
        if sex=1 **then**
            Sexual reproduction, sample n_r and $\boldsymbol{L_r}$ (see II) from j
        end if
    end for

45: **end if**
46: **C**
47: Selection, genetic drift, asexual reproduction at time t:
48: See **A**

Algorithm 4 ABC-rejection algorithm for summary statistics calculated from datasets simulated from the **Simulator** and their corresponding parameters from the prior distribution. Input: proportion of simulations to accept tol, number of simulated data sets $nsim$, number of summary statistics per simulation $nstat$.

1: Input: tol, $nsim$, $nstat$.
2: Input: $\boldsymbol{\theta}_i^*$. ▷ Simulated from the prior $\boldsymbol{\theta}_i^* \sim P(\boldsymbol{s}, m, sex, \boldsymbol{l}; \boldsymbol{K})$ in the Simulator.
3: Sample $\boldsymbol{x}_i^* \sim P(\boldsymbol{x}|\boldsymbol{s}, m, sex, \boldsymbol{l}; \boldsymbol{K})$. ▷ Sampled from the Simulator.
4: Calculate $\boldsymbol{SumStat}_i^* = S(\boldsymbol{x}_i^*)$.
5: Calculate $d_i = \sqrt{\sum_{j=1}^{nstat} \left(\frac{SumStat_{i,j}^* - SumStat_{obs,j}}{SD_{,j}} \right)^2}$.
6: Accept order $\left(d_1, d_2, \cdots, d_{nsim}\right)[1 : (nsim \cdot tol)]$ of $\|.\| = M = (nsim \cdot tol)$. ▷ Indeces of accepted ordered M values denoted by $id_{1:M}$.
7: Return mth accepted $(\boldsymbol{\theta}_{\boldsymbol{m}}^*, \boldsymbol{SumStat}_{\boldsymbol{m\ scaled}}^*)$. ▷ $\boldsymbol{SumStat}_{\boldsymbol{m\ scaled}}^* = \left(\frac{SumStat_{m,1}^*}{SD_{,1}}, \frac{SumStat_{m,2}^*}{SD_{,2}}, \cdots, \frac{SumStat_{m,nstat}^*}{SD_{,nstat}} \right)$.

Algorithm 5 ABC-linear regression correction algorithm performed on the output of the ABC-rejection. Input: $\boldsymbol{SumStat}_{\boldsymbol{obs\ scaled}}$, mth accepted $\boldsymbol{SumStat}_{\boldsymbol{m\ scaled}}^*$, mth accepted $\boldsymbol{\theta}_{\boldsymbol{m}}^*$ from ABC-rejection, with indeces of accepted ordered M values denoted by $id_{1:M}$.

1: Input: $\boldsymbol{SumStat}_{\boldsymbol{obs\ scaled}}$, $\boldsymbol{SumStat}_{\boldsymbol{id_{1:M}\ scaled}}^*$, $\boldsymbol{\theta}_{\boldsymbol{id_{1:M}}}^*$.
2: Calculate each weight $w_m = \frac{\kappa_{d_M}(d_m)}{\sum_{m=1}^{M} \kappa_{d_M}(d_m)}$. ▷ Calculate each of M weights, $\{w_1, w_2, \ldots, w_M\}$, based on kernel type.
3: Calculate $\hat{\boldsymbol{\beta}}_{WLS} = (\langle \mathbf{1}, \boldsymbol{SumStat}_{\boldsymbol{obs\ scaled}} \rangle^\top \boldsymbol{W}\, \boldsymbol{SumStat}_{\boldsymbol{obs\ scaled}})^{-1} \langle \mathbf{1}, \boldsymbol{SumStat}_{\boldsymbol{obs\ scaled}} \rangle^\top \boldsymbol{W}\, \boldsymbol{\theta}_{\boldsymbol{id_{1:M}}}^*$. ▷ WLS estimate of $\boldsymbol{\beta}$ from *lsfit()*, where $\mathrm{diag}(\boldsymbol{W}) = \{w_1, w_2, \ldots, w_M\}$.
4: Return adjusted values $\boldsymbol{\theta}_{\boldsymbol{id_{1:M}\ adj}}^* = \boldsymbol{\theta}_{\boldsymbol{id_{1:M}}}^* - (\left\langle \mathbf{1}, \boldsymbol{SumStat}_{\boldsymbol{id_{1:M}\ scaled}}^* \right\rangle - \langle \mathbf{1}, \boldsymbol{SumStat}_{\boldsymbol{obs\ scaled}} \rangle) \hat{\boldsymbol{\beta}}_{WLS}$.

Table 3
Description of conversion between genomic recombination rate, and recombination rate with respect to SNPs.

One crossover of actual genome that occurs between 2 consecutive SNPs actually would be one crossover between simulated SNPs only. Two crossovers between consecutive SNPs on the genome would look like no crossover between simulated SNPs, which we do not want to consider because it is very rare. Below is a note of how the recombination conversion is performed:

For yeast cross from the lab as example, genome length=12.066Mb and strains differ by 73,015 SNPs
$SNP_{spacing} = 12.066\text{Mb}/73,015 \approx 165$
The recombination rate conversion inside the function:

$avg_{numrec} = (L \times SNP_{spacing} - 1) \times r$ (This is the average number of recombination events of genome chunk on which the SNPs are spaced)

Example: For $L = 100$, and $r = 3.0 \times 10^{-4}$ $avg_{numrec} = 4.95$ Convert r: $r_{SNP} = avg_{numrec}/(L-1) = 4.95/99 = 0.05$ (This is the recombination rate of SNPs alone) $avg_{numrecSNP} = (L-1) \times r_{SNP} = 99 \times 0.05 = 4.95 = avg_{numrec}$ (This is the average number of recombination events of SNPs alone)

Parameters per SNP equivalent to line 13 and line 15 of Algorithm 3:
Sample number of recombination events independently of each other (line 13):
$n_r \sim \text{Bin}(L-1, r_{SNP})$

Sample loci of recombination events (line 15): $\boldsymbol{L_{rSNP}} \sim \text{Discrete Unif}[n_r, L-1]$

What is written in line 13 and line 15 of Algorithm 3:
Sample number of recombination events independently of each other (line 13):
$n_r \sim \text{Bin}(L \times SNP_{spacing} - 1, r)$

Sample loci of recombination events (line 15):
$\boldsymbol{L_r} \sim \text{Discrete Unif}[n_r, L \times SNP_{spacing} - 1]$

The parents reproduce sexually and parental genome recombination takes place. We denote probability of genome recombination with r (Morgans/bp units). We sample number of recombination events n_r from Binomial distribution with a sequence of independent $L \times SNP_{spacing} - 1$ experiments with probability r. For each of the n_r recombination events, we sample genome loci of recombination events $\boldsymbol{L_r}$, from Discrete Uniform distribution from locus 1 to locus $L \times SNP_{spacing} - 1$, but we assume no more than one recombination event between two consecutive SNPs. The cross of two populations at $t = 0$ (end Part II and start Part A of Algorithm 3) serves as an ancestral pool for $\boldsymbol{X}$ and $\boldsymbol{Y}$ populations.

Table 4
Euclidean distance equations.

Euclidean distance:
In the ABC-rejection we apply the Euclidean distance [39,40] technique to scale summary statistics:

$$d_i = \sqrt{\sum_{j=1}^{nstat} \left(\frac{SumStat_{i,j} - SumStatat_{obs,j}}{SD_j} \right)^2} \tag{9}$$

with $SD, j,\ j = 1, 2, \ldots, nstat$, of the median absolute deviation. The $SumStat_{i,j}$ and $SumStat_{obs,j}$ are points in $nstat$-dimensional space.
We scale the $nsim \times nstat$ summary statistics matrix such that each of the summary statistics contributes equally when SD on the $nstat$-dimensional space:

$$\begin{bmatrix} \frac{SumStat^*_{1,1}}{k\cdot\text{median}\left(|SumStat^*_{i,1} - \text{median}(\boldsymbol{SumStat}^*_{,1})|\right)} & \cdots & \frac{SumStat^*_{1,nstat}}{k\cdot\text{median}\left(|SumStat^*_{i,nstat} - \text{median}(\boldsymbol{SumStat}^*_{,nstat})|\right)} \\ \vdots & \ddots & \vdots \\ \frac{SumStat^*_{nsim,1}}{k\cdot\text{median}\left(|SumStat^*_{i,1} - \text{median}(\boldsymbol{SumStat}^*_{,1})|\right)} & \cdots & \frac{SumStat^*_{nsim,nstat}}{k\cdot\text{median}\left(|SumStat^*_{i,nstat} - \text{median}(\boldsymbol{SumStat}^*_{,nstat})|\right)} \end{bmatrix}$$
$$\equiv \begin{bmatrix} \frac{SumStat^*_{1,1}}{SD_{,1}} & \cdots & \frac{SumStat^*_{1,nstat}}{SD_{,nstat}} \\ \vdots & \ddots & \vdots \\ \frac{SumStat^*_{nsim,1}}{SD_{,1}} & \cdots & \frac{SumStat^*_{nsim,nstat}}{SD_{,nstat}} \end{bmatrix}, \tag{10}$$

where k is a constant scale factor dependent of the distribution of summary statistics [59] such that the median absolute deviation is a consistent estimator for the standard deviation of each summary statistic for $j = 1, 2, \ldots, nstat$:

$$k \cdot \mathbb{E}[\text{median}\left(|SumStat_{i,j} - \text{median}(\boldsymbol{SumStat}_{,j})|\right)] = \sigma_j. \tag{11}$$

For Gaussian distribution $k = 1/\Phi^{-1}(3/4) \approx 1.4826$ [59] (estimated after converting to standard normal), where for uniform distribution $k = 2/\sqrt{3}$ (for standard continuous uniform which spans on $[-\sqrt{3}, \sqrt{3}]$, with $k \cdot (\sqrt{3}/2) = 1$). All of the parameters are simulated from uniform priors, therefore we used $k = 2/\sqrt{3}$.

Table 5
Kernel equations.

Kernels and linear regression correction:
Associated with each accepted pair $(\boldsymbol{\theta}^*_{\boldsymbol{m}}, \boldsymbol{SumStat}^*_{\boldsymbol{m\ scaled}})$ from Algorithm 1 output a weight:

$$w_m \propto \kappa_{d_M}(d_m) \tag{12}$$

[43], and

$$w_m = \frac{\kappa_{d_M}(d_m)}{\sum_{m=1}^{M} \kappa_{d_M}(d_m)} \tag{13}$$

[41], such that $\sum_{m=1}^{M} w_m = 1$, to obtain a weighted sample $(\boldsymbol{\theta}^*_{\boldsymbol{m\ adj}}, w_m)$. The w_m can be calculated based on either Epanechnikov kernel:

$$\kappa_{d_M}(d_m) = \left(\frac{3}{4}\right)\left[1 - \left(\frac{d_m}{d_M}\right)^2\right] \tag{14}$$

[45], or the Gaussian kernel:

$$\kappa_{d_M}(d_m) = \left(\frac{1}{\sqrt{2\pi}}\right)\exp\left[-\left(\frac{1}{2}\right)\left(\frac{d_m}{d_M}\right)^2\right] \tag{15}$$

[41,44]. Next, we implemented *lsfit* R function [60,61] to calculate weighted least squares estimates of regression coefficients, with indeces of accepted ordered M values denoted by $id_{1:M}$:

$$\hat{\boldsymbol{\beta}}_{WLS} = (\langle \mathbf{1}, \boldsymbol{SumStat}_{obs\ scaled}\rangle^{\top} \boldsymbol{W}\, \boldsymbol{SumStat}_{obs\ scaled})^{-1} \langle \mathbf{1}, \boldsymbol{SumStat}_{obs\ scaled}\rangle^{\top} \boldsymbol{W}\, \boldsymbol{\theta}^*_{\boldsymbol{id}_{1:M}} \tag{16}$$

,where $\boldsymbol{W}$ is a diagonal matrix, and $\mathrm{diag}(\boldsymbol{W}) = \{w_1, w_2, \ldots, w_M\}$ corresponding to kernel weights. The beta estimates vector is an approximate draw from the posterior with $\boldsymbol{SumStat}^*_{\boldsymbol{id}_{1:M}\ scaled} = \boldsymbol{SumStat}_{obs\ scaled}$, and parameters from regression correction $\boldsymbol{\theta}^*_{\boldsymbol{id}_{1:M}}$ are corrected proportional to the $(\left\langle \mathbf{1}, \boldsymbol{SumStat}^*_{\boldsymbol{id}_{1:M}\ scaled} \right\rangle - \langle \mathbf{1}, \boldsymbol{SumStat}_{obs\ scaled}\rangle)\hat{\boldsymbol{\beta}}_{WLS}$:

$$\boldsymbol{\theta}^*_{\boldsymbol{id}_{1:M\ adj}} = \boldsymbol{\theta}^*_{\boldsymbol{id}_{1:M}} - (\left\langle \mathbf{1}, \boldsymbol{SumStat}^*_{\boldsymbol{id}_{1:M\ scaled}} \right\rangle - \langle \mathbf{1}, \boldsymbol{SumStat}_{obs\ scaled}\rangle)\hat{\boldsymbol{\beta}}_{WLS} \tag{17}$$

[4,62,63], such that accepted data set with smallest Euclidean distance d_i from Algorithm 1 output is adjusted the least and the accepted data set with largest d_i out of the M accepted is adjusted the most.

Appendix B

Detailed description of the assessment of genetic drift on the model from Section 4.1.

In order to assess the effect of genetic drift and most informative sample summary statistic, we did the following in order to compare mean square errors (MSEs) of s: (1) we examined change in difference of allele frequency between population $\boldsymbol{X}$ and $\boldsymbol{Y}$ for every generation of the model for a single locus under selection (midpoint SNP), with reduced the model to $L = 20$ SNPs, with selection at locus at $(L/2)$th SNP rounded up of $s_{L/2} = 0.1$, migration $m = 0.1$, spacing of SNPs every 165 loci ($SNP_{spacing} = 165$), and recombination rate $r = 2.880 \times 10^{-4}$; (2) we compared the stochasticity of a single randomly chosen simulation to an average of the change in allele frequencies per SNP from $nsim = 100,000$ simulations; (3) we simulated a single autosomal locus (a deterministic case, infinite population size) [38] under same strength of selection, with equal proportions of allele 0 and 1 (expected values from the simulated model) – both 0.5 – in both populations at generation $t = 0$ (equivalent to $t = 0$ in **Part A** in Fig. 1), and we updated frequencies per generation to:

$$\text{frequency after selection} \propto \text{newborn frequency} \times \text{viability}, \tag{9}$$

We show negligent effect of genetic drift and effective population size of simulations on the difference in allele frequency p between population $\boldsymbol{X}$ and population $\boldsymbol{Y}$, $(p_X - p_Y)$, summary statistics at locus under selection when we average our summary statistics at each SNP over 100,000 simulations. By removing the stochastic effect we can deterministically identify most informative summary statistic via ABC approach. Here we track $(p_X - p_Y)$ per generation based on the model.

Fig. 10 (main, left plot) shows mean of $(p_X - p_Y)$ summary statistics of 100,000 simulations tracked from $t = 0$ to generation at end of the simulation. There, we see linkage disequilibrium between locus under selection (red line) and the neighboring loci (black lines), proportional to the distance between locus under selection and those that are not. The three drops in $(p_X - p_Y)$ summary statistics values correspond to the migration generations where the populations mixing occurs and therefore less difference seen in the proportion of alleles between populations $\boldsymbol{X}$ and $\boldsymbol{Y}$. In the upper right plot we superimposed the red plot from the main, left plot and deterministically derived $(p_X - p_Y)$ for a single locus case [38] for haploid case with applied migration every 50th generation in green. We see that we cannot distinguish between single locus deterministically derived $(p_X - p_Y)$ values, and the red line (locus under selection, main, left plot) of the mean of $(p_X - p_Y)$ summary statistics of 100,000 simulations. For reference how stochastic a single simulation can be, we randomly picked one of the 100,000 simulations and plotted the $(p_X - p_Y)$ values (lower right plot). This visual representation verifies that with very large number of simulations the values of summary statistics $(p_X - p_Y)$ converge to deterministic (expected) values, and the signal of selection has an effect on the nearby loci (linkage disequilibrium effect).

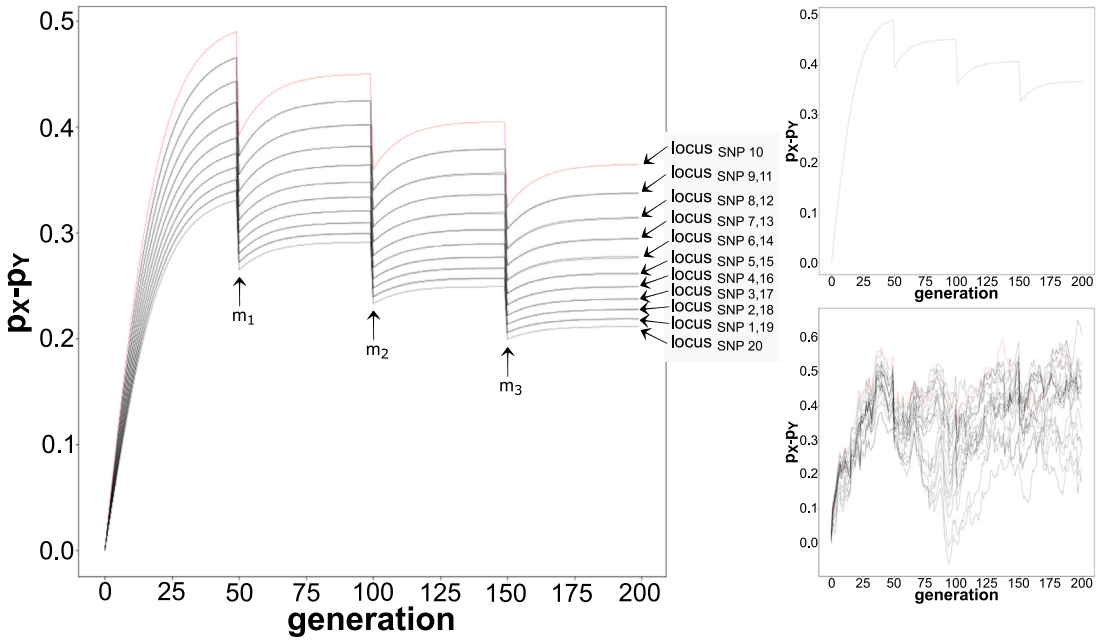

Fig. 10. Mean of summary statistic $(p_X - p_Y)$ values per locus at SNP 1 to $L = 20$ for $nsim = 100,000$ simulated data sets, with selection at single locus at SNP 10. The decrease in $(p_X - p_Y)$ is shown to be proportional to distance away from locus under selection, and with drops of $(p_X - p_Y)$ values at migration m generations (left). Superimposed expected values (deterministic model), and mean of 100,000 simulated data sets of $(p_X - p_Y)$ at locus under selection (upper right). Visible genetic drift for a randomly chosen simulated data set.

Appendix C. Supplementary figures

Supplementary figures for the 1st method, i.e. without the initial outlier scan.
Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jcmds.2024.100091.

References

[1] Butlin R, Debelle A, Kerth C, Snook RR, Beukeboom LW, Cajas RC, et al. What do we need to know about speciation? Trends Ecol Evol 2012;27(1):27–39.
[2] Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 2008;454(7203):479–85.
[3] Mitchell AP. Control of meiotic gene expression in saccharomyces cerevisiae. Microbiol Mol Biol Rev 1994;58(1):56–70.
[4] Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics 2002;162(4):2025–35.
[5] Barton NH. Genetic hitchhiking. Philos Trans R Soc London Ser B: Biol Sci 2000;355(1403):1553–62.
[6] Nosil P, Feder JL. Genomic divergence during speciation: causes and consequences. Philos Trans R Soc B 2012;367(1587):332–42.
[7] Qanbari S. On the extent of linkage disequilibrium in the genome of farm animals. Front Genet 2020;10:1304.
[8] Hether TD. Genetic networks, adaptation, & the evolution of genomic islands of divergence. University of Idaho; 2016.
[9] Ruderfer DM, Pratt SC, Seidel HS, Kruglyak L. Population genomic analysis of outcrossing and recombination in yeast. Nature Genet 2006;38(9):1077–81.
[10] Illingworth CJ, Parts L, Bergström A, Liti G, Mustonen V. Inferring genome-wide recombination landscapes from advanced intercross lines: application to yeast crosses. PLoS One 2013;8(5):e62266.
[11] Bast J, Jaron KS, Schuseil D, Roze D, Schwander T. Asexual reproduction reduces transposable element load in experimental yeast populations. Elife 2019;8:e48548.
[12] Hamilton G, Currat M, Ray N, Heckel G, Beaumont M, Excoffier L. Bayesian estimation of recent migration rates after a spatial expansion. Genetics 2005;170(1):409–17.
[13] Boitard S, Rodriguez W, Jay F, Mona S, Austerlitz F. Inferring population size history from large samples of genome-wide molecular data-an approximate Bayesian computation approach. PLoS Genet 2016;12(3):e1005877.
[14] Tsai IJ, Burt A, Koufopanou V. Conservation of recombination hotspots in yeast. Proc Natl Acad Sci 2010;107(17):7847–52.
[15] Hermann P, Heissl A, Tiemann-Boege I, Futschik A. Ldjump: Estimating variable recombination rates from population genetic data. Mol Ecol Resour 2019;19(3):623–38.
[16] Gallet R, Cooper TF, Elena SF, Lenormand T. Measuring selection coefficients below 10- 3: method, questions, and prospects. Genetics 2012;190(1):175–86.
[17] Tallmon DA, Koyuk A, Luikart G, Beaumont MA. Computer programs: onesamp: A program to estimate effective population size using approximate Bayesian computation. Mol Ecol Resour 2008;8(2):299–301.
[18] Li S, Jakobsson M. Estimating demographic parameters from large-scale population genomic data using approximate Bayesian computation. BMC Genet 2012;13(1):22.
[19] Quinto-Cortés CD, Woerner AE, Watkins JC, Hammer MF. Modeling SNP array ascertainment with approximate Bayesian computation for demographic inference. Sci Rep 2018;8(1):1–10.

[20] Smith CC, Flaxman SM. Leveraging whole genome sequencing data for demographic inference with approximate bayesian computation. Mol Ecol Resour 2020;20(1):125–39.
[21] Estoup A, Beaumont M, Sennedot F, Moritz C, Cornuet J-M. Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, Bufo marinus. Evolution 2004;58(9):2021–36.
[22] Fagundes NJ, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, et al. Statistical evaluation of alternative models of human evolution. Proc Natl Acad Sci 2007;104(45):17614–9.
[23] Cornuet J-M, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, et al. DIYABC v2. 0: A software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. Bioinformatics 2014;30(8):1187–9.
[24] Sousa V, Beaumont M, Fernandes P, Coelho M, Chikhi L. Population divergence with or without admixture: selecting models using an ABC approach. Heredity 2012;108(5):521–30.
[25] Wright S. The genetical structure of populations. Ann Eugenics 1949;15(1):323–54.
[26] Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F ST. Nature Rev Genet 2009;10(9):639–50.
[27] Beaumont MA, Nichols RA. Evaluating loci for use in the genetic analysis of population structure. Proc R Soc B 1996;263(1377):1619–26.
[28] Laayouni H, Montanucci L, Sikora M, Mele M, Dall'Olio GM, Lorente-Galdos B, et al. Similarity in recombination rate estimates highly correlates with genetic differentiation in humans. PLoS One 2011;6(3):e17913.
[29] Kimura M. The neutral theory of molecular evolution. Cambridge University Press; 1983.
[30] Otyama PI, Wilkey A, Kulkarni R, Assefa T, Chu Y, Clevenger J, et al. Evaluation of linkage disequilibrium, population structure, and genetic diversity in the US peanut mini core collection. BMC Genomics 2019;20(1):1–17.
[31] Bomba L, Nicolazzi EL, Milanesi M, Negrini R, Mancini G, Biscarini F, et al. Relative extended haplotype homozygosity signals across breeds reveal dairy and beef specific signatures of selection. Genet Select Evol 2015;47(1):1–14.
[32] Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature 2002;419(6909):832–7.
[33] Weigand H, Leese F. Detecting signatures of positive selection in non-model species using genomic data. Zool J Linnean Soc 2018;184(2):528–83.
[34] Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature 2007;449(7164):913–8.
[35] Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol 2006;4(3):e72.
[36] Wagh K, Bhatia A, Alexe G, Reddy A, Ravikumar V, Seiler M, et al. Lactase persistence and lipid pathway selection in the Maasai. Public Library of Science San Francisco, USA; 2012.
[37] Qanbari S, Gianola D, Hayes B, Schenkel F, Miller S, Moore S, et al. Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. BMC Genomics 2011;12(1):1–12.
[38] Gillespie JH. Population genetics: A concise guide. JHU Press; 2004.
[39] Greenacre M. Correspondence analysis in practice. CRC Press; 2017.
[40] Prangle D, et al. Adapting the ABC distance function. Bayesian Anal 2017;12(1):289–309.
[41] Park M, Jitkrittum W, Sejdinovic D. K2-ABC: Approximate Bayesian computation with kernel embeddings. In: Artificial intelligence and statistics. PMLR; 2016, p. 398–407.
[42] Nott DJ, Fan Y, Marshall L, Sisson S. Approximate Bayesian computation and Bayes' linear analysis: toward high-dimensional ABC. J Comput Graph Statist 2014;23(1):65–86.
[43] Blum MG. Regression approaches for approximate bayesian computation. 2017, arXiv preprint arXiv:1707.01254.
[44] Csilléry K, Blum MG, Gaggiotti OE, François O. Approximate Bayesian computation (ABC) in practice. Trends Ecol Evol 2010;25(7):410–8.
[45] Epanechnikov VA. Non-parametric estimation of a multivariate probability density. Theory Probab Appl 1969;14(1):153–8.
[46] Duan S-F, Han P-J, Wang Q-M, Liu W-Q, Shi J-Y, Li K, et al. The origin and adaptive evolution of domesticated populations of yeast from far East Asia. Nature Commun 2018;9(1):1–13.
[47] Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, et al. Popoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 2011;6(1):e15925.
[48] Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS genet 2010;6(2):e1000862.
[49] Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Res 2009;19(5):826–37.
[50] Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet 2013;47:97–120.
[51] Gallone B, Steensels J, Prahl T, Soriaga L, Saels V, Herrera-Malaver B, et al. Domestication and divergence of saccharomyces cerevisiae beer yeasts. Cell 2016;166(6):1397–410.
[52] Wolf JB, Ellegren H. Making sense of genomic islands of differentiation in light of speciation. Nature Rev Genet 2017;18(2):87–100.
[53] Alachiotis N, Pavlidis P. Scalable linkage-disequilibrium-based selective sweep detection: A performance guide. GigaScience 2016;5(1):s13742–016.
[54] Wang Z, Wang L, Szolnoki A, Perc M. Evolutionary games on multilayer networks: A colloquium. Eur Phys J B 2015;88:1–15.
[55] Mariac C, Ousseini IS, Alio A-K, Jugdé H, Pham J-L, Bezançon G, et al. Spatial and temporal variation in selection of genes associated with pearl millet varietal quantitative traits in situ. Front Genet 2016;7:130.
[56] Slatkin M. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. Nature Rev Genet 2008;9(6):477–85.
[57] Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res 1966;8(3):269–94.
[58] Comeron JM, Williford A, Kliman R. The Hill–Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity 2008;100(1):19–31.
[59] Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. J Amer Statist Assoc 1993;88(424):1273–83.
[60] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021, URL https://www.R-project.org/.
[61] Becker RA, Chambers JM, Wilks AR. The new S language: A programming environment for data analysis and graphics. Wadsworth and Brooks/Cole Advanced Books & Software; 1988.
[62] Rodrigues G, Nott DJ, Sisson SA. Functional regression approximate Bayesian computation for Gaussian process density estimation. Comput Statist Data Anal 2016;103:229–41.
[63] Blum MG, François O. Non-linear regression models for approximate Bayesian computation. Stat Comput 2010;20(1):63–73.