A statistical test for lineage-specific natural selection on quantitative traits based on multiple-line crosses

N. Riedel[1], B. S. Khatri[2], M. Lässig[1], and J. Berg[1]
Institut für Theoretische Physik, University of Cologne - Zülpicher Straße 77, 50937 Köln, Germany
[2]*Mathematical Biology Division, National Institute for Medical Research, The Ridgeway, London, U.K.*

Phenotypic differences between species may be attributable to natural selection. However, it is a difficult task to quantify the strength of evidence for selection acting on a particular trait. Here we develop a population-genetic test for selection acting on a quantitative trait, which is based on multiple-line crosses. We show that using multiple lines increases both the power and the scope of selection inference. First, a test based on three or more lines detects selection on a quantitative trait with strongly increased statistical significance, which is quantified by our analysis. Second, a multiple-line test allows to distinguish selection from neutral evolution as well as lineage-specific selection from selection under uniform selection strength. This is in contrast to tests based on two lines, where only differences in selection coefficients can be inferred. Our analytical results are complemented by extensive numerical simulations. We apply the multiple-line test to QTL data on floral character traits in plant species of the *Mimulus* genus and on photoperiodic traits in different maize strains. In both cases, we find a signature of lineage-specific selection that is not seen in a two-line test. We also extend the multiple-line test to short divergence times.

I. INTRODUCTION

Extensive experimental work has helped reveal the genetic architecture of quantitative traits [1–6], allowing to study the basis of trait variation within and across species. A long-term goal of QTL research is to understand the mapping from genotype to phenotype underlying a particular quantitative trait. Crosses between individuals from different lines are used to identify loci whose allelic states are statistically correlated with a particular trait. However, the ability of QTL studies to identify the molecular basis of quantitative traits is still limited; it is especially difficult to pinpoint genetic loci influencing a trait [7]. Targeted efforts have been made to resolve loci at the level of single genes or even nucleotides [8–13], but these cases are still the exception.

In recent years, the QTL experiments were also extended to crosses between multiple lines. Harnessing information from several lines drastically increases the power and accuracy of QTL identification [14, 15]), allowing to test for epistatic interactions [16, 17], and increasing the genetic variability that can be accessed [16]. For instance, all loci that have the same allele in two lines also have the same allele in all crosses of these lines. In the absence of genetic variance, the effect of such a locus on a trait cannot be determined. Analysing more than two lines increases the number of loci that differ in allelic state in at least one line, allowing to identify more loci affecting a quantitative trait. Multiple-line pairwise crosses are most common in animal and plant breeding [16, 18], where often many different lines are available for crossing. However, the extension to multiple line crosses also brings new challenges. For instance, choosing the right mating design for the QTL experiments is important for multiple-line crosses [19, 20]. Since most statistical methods for QTL identification developed for two-line crosses cannot easily be extended to the multiple-line case, new and more sophisticated methods were developed [21]. These methods are based on least-squares regression [22], maximum likelihood [21, 23], and a Bayesian approach [24] and have been applied to a range of experimental datasets [15, 16, 18, 25, 26].

Beyond the identification of QTL, an important aim of QTL analysis is to infer the evolutionary forces acting on a particular trait. Here, the central question is whether and how natural selection acted on a trait during its evolutionary history. A more specific question is whether the strength of selection is constant across a phylogeny, or whether it acted in a lineage-specific manner. Several statistical tests make use of the data gained from QTL experiments to detect effects of natural selection. The test of Orr [27] asks if the statistical distribution of alleles differs from the statistics expected under neutral evolution: Orr's test statistics looks at the excess of alleles that increase the value of the trait ("+" alleles) in a line, and at the distribution of trait contributions of these alleles. The test of Rice and Townsend [28] combines QTL analysis with data from mutation accumulation experiments and asks if mutations seen between two lines tend to affect the trait more than those seen in experiments that accumulate largely neutral mutations. The test of Fraiser [29, 30] focuses on the inference of selection on expression values by expression QTL (eQTL). It distinguishes lineage-specific positive selection from relaxed negative selection by means of expression data from a third lineage as outgroup. However, no test to date uses the full statistical information from multiple-line QTL experiments.

In this paper, we develop a statistical framework to test different evolutionary hypotheses for multiple QTL lines against each other. Using a systematic log-likelihood scoring, we find that a multiple-line test allows identification of selection with increased statistical significance compared to the two line test. However, the consequences of multiple-line testing go beyond the mere increase of the number of observed loci. For two lines, it turns out that only relative differences in selection strength can be inferred when the allele statistics have reached a steady state (and only a weak signal on absolute selection strength is present before the steady state is reached). A qualitatively different situation arises if QTL information is available for three or more lines, which allows inference of the absolute selection strength of all lines. We find a similar situation also in the context of corrections for multiple testing when the trait under investigation is chosen from an unknown trait pool.

In the following, we develop a log-likelihood score that quantifies the likelihood of neutral and different selective hypotheses in an explicit evolutionary framework. We first explore our approach on artificial data and then apply our test to floral quantitative traits in different *Mimulus* species and to photoperiod traits in maize.

II. A GENERAL n-LINE SELECTION MODEL

Central to our analysis is a quantitative trait G determined by k loci labelled $i = 1, \ldots, k$. For n lines one can have n different alleles at each locus. Here we restrict ourselves to the biallelic case, assuming only two possible alleles $q_{a,i} = \pm 1$ at locus i for line $a = 1 \ldots n$. This approximation can be understood as a restriction to the mutation with the largest effect on the locus. This assumption is supported by experimental studies on crosses between 4 different lines, where the majority of loci show only two significantly different alleles [16, 26]. In a study using crosses between 25 lines, however, many alleles with different trait contributions have been observed at the same locus [31]. Nevertheless, for crosses between only few lines, the two-allele assumption should be reasonable, since the uncertainty of the estimates of the trait contributions make it hard to distinguish different alleles with similar trait contributions.

We assume a linear trait model without trait epistasis; the allelic state at each locus contributes additively to the trait $G = \sum_{i=1}^{k} g_i q_i$, where an additive trait contribution g_i comes from locus i. Without loss of generality we take $g_i \geq 0$, so $q_i = +1$ (termed the +-allele) results in a higher trait value than $q_i = -1$ (the −-allele). We note that even if selection on the trait is quadratic [32–35], the selection difference between lines remains linear.

We now consider a haploid population in the weak-mutation regime with full recombination. We assume a linear fitness landscape $F = s \sum_{i=1}^{n} g_i q_i$ with constant selection strength $s > 0$, resulting in a selection coefficient $\sigma_i = N s g_i$ for each locus proportional to the trait contribution [36, 37]. The assumption of a linear fitness landscape implies that the allelic states at different loci are statistically independent, so the allele statistics factorizes over loci. Due to the factorization it is sufficient to consider the allele statistics at a single locus with given trait contribution g; the locus index i will be omitted in the following where possible.

First, we consider the limit of long evolutionary times between lines, where the allele statistics at each locus no longer change with time. Later, we will discuss the validity of this assumption in more detail and compare with the complementary regime of short evolutionary times.

We now evaluate the allele statistics of the loci in a scenario where the quantitative trait G is under selection. In the most general case, each line a evolves under a different selection strength s_a. For line a under selection with selection strength s_a, Kimura-Ohta theory [38] for finite populations evolving under genetic drift and selection gives an equilibrium distribution [39–41]

$$P(q_a|g) = \frac{e^{N s_a g q_a}}{e^{N s_a g} + e^{-N s_a g}} \tag{1}$$

for the probability of allele $q_a = \pm 1$ at a locus with trait effect g. Here, N is the effective population size.

If the n lines have evolved independently for a long time, the joint probability distribution for all lines factorizes, giving $P(q_1, q_2, \ldots, q_n|g) = \prod_{a=1}^{n} P(q_a|g)$. Crucially, in QTL analysis based on crosses between individuals from different lines only the effects of loci differing in state between at least two lines can be determined. For this reason, the two states with $q_1 = q_2 = \ldots = q_n = \pm 1$ remain unobserved. In the following, we denote the number of diverged loci k_{div}.

The joint allele distribution for a given locus follows as

$$P(q_1, \ldots, q_n|g) = \frac{e^{g \sum_{i=1}^{n} q_i N s_i}}{\sum\limits_{\{q_1', q_2', \ldots, q_n' = \pm 1\}} e^{g \sum_{i=1}^{n} q_i' N s_i}}\,, \tag{2}$$

for all possible states of $q_1, q_2, \ldots, q_n$ excluding the two unobserved states with $q_1 = q_2 = \ldots = q_n$. For each line i the probability to have the +-allele increases with increasing selection strength s_i (or decreases for negative s_i).

III. LOG-ODDS SCORING OF DIFFERENT EVOLUTIONARY SCENARIOS

Based on the above result, different restricted selection scenarios can be considered, e.g. neutral evolution with $s_1 = s_2 = \ldots = s_n = 0$ or an equal selection strength in all lines with $s_1 = s_2 = \ldots = s_n = s$. For a given experimental dataset it is straightforward to estimate the selection strengths s_i for all lines by maximizing the likelihood (2) with respect to the selection strengths. However, when only few loci are known for a trait, the inference of all selection strengths may be unreliable due to overfitting. In this case it is convenient to restrict the parameter space of selection strengths and to test specific hypotheses against each other. For example, one can compare a neutral scenario ($s_1 = s_2 = \ldots = s_n = 0$), a scenario with uniform selection strength on all lines ($s_1 = s_2 = \ldots = s_n = s$), or lineage-specific selection patterns ($s_1 \neq s_2 = \ldots = s_n = s$). We use log-odds ratios to weight the evidence for different evolutionary scenarios against each other. The allele statistics under two evolutionary scenarios P and Q are then compared to each other using the log-odds score

$$S_{Q,P} = \sum_{i=1}^{k} \ln\left(\frac{Q(q_{1,i}, q_{2,i}, \ldots, q_{n,i}|g_i)}{P(q_{1,i}, q_{2,i}, \ldots, q_{n,i}|g_i)}\right)\,. \tag{3}$$

Here, the restricted selection parameters of both scenarios are determined via maximum-likelihood. For the Q-scenario the parameters are chosen to maximize the score $S_{Q,P}$ and for the P-scenario to minimize $S_{Q,P}$. This score quantifies the evidence for or against particular evolutionary scenarios given allelic states of different loci and their contributions to the trait. The score (3) is positive if loci follow an allele statistics more in agreement with the allele distribution of scenario Q than with the allele distribution of scenario P.

If two scenarios with different numbers of free parameters are tested against each other, the score (3) will be biased towards the scenario with more parameters. A simple way to correct this bias is the Bayesian information criterion

(BIC) [42]. The score (3) is then decreased by an offset $S_{Q,P} = \ln(Q/P) - l/2 \ln k$ where l is the excess number of parameters in model Q and k is the number of loci.

IV. INCREASED POWER IN MORE THAN TWO LINES

There is a simple reason why the power of the selection test increases when more lines are used. Since only loci that are diverged between the lines can be observed, a certain fraction of loci remain hidden. For two lines, loci with the states $(q_1, q_2) = (++)$ and $(--)$ cannot be observed, which account on average for 50% of the loci under neutrality. For three lines we only have 2 unobserved out of 8 possible configurations, decreasing the average fraction of unobserved loci to 25%. The fraction of unobserved loci decreases with the number of lines.

To probe the log-odds score (3) for a varying number of lines, we perform numerical simulations to test a selective and a neutral hypotheses against each other on artificial data. For $n = 1 \ldots 6$ lines and a number of $k = 20$ loci we draw trait contributions g_i randomly from a gamma distribution [27, 43]. Then we simulate the evolutionary dynamics of these loci under different evolutionary scenarios, which we label for easy reference. In the first, neutral scenario P_0, the selection strength acting on all lines is zero ($s_1 = s_2 = \ldots = s_n = 0$). In the second scenario Q_1 only line 1 is under selection ($s_1 = s$, $s_2 = \ldots = 0$).

We begin by simulating the evolutionary dynamics of k loci under scenario Q_1. Each run results in a set of alleles $\{q_{1,i}, \ldots, q_{n,i}\}$ at the different loci. For the subset of diverged loci we compute the log-odds score S_{Q_1,P_0} (3). This score quantifies how likely scenario Q_1 is relative to scenario P_0. To gauge the statistical significance of a given value of this score, we also estimate the probability of reaching the same score or higher under scenario P_0. This p-value measures the rate of false positives, that is the fraction of realizations where a certain score favouring scenario Q_1 is reached under the statistics of scenario P_0. We compute this p-value by performing a large number of runs under the dynamics of P_0 to see what fraction of them gave a score matching or exceeding S. To gauge how frequently a positive score in favour of scenario Q_1 in line 1, 2 or 3 occurs, we sort the configurations drawn from the null model P_0 according to their phenotypes G_i. This typically reduces the significance of the test compared to more specific questions as, e.g. how likely is it to observe given score in favour of lineage-specific selection in line 1 specifically.

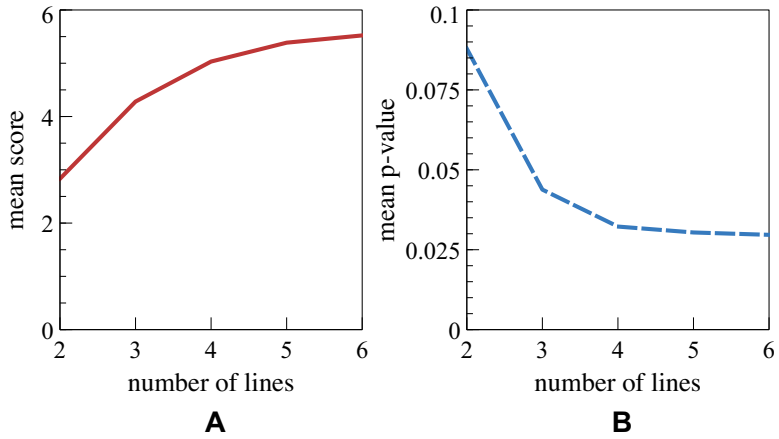

FIG. 1. **Log-odds score and statistical significance of the test for a fixed number of loci.** (A) The log-odds score (3) averaged over many realizations of the allele statistics under the selective scenario Q_1 tested against the neutral scenario P_0 for different numbers of lines at a fixed total number of $k = 20$ loci. The score increases with the number of lines as on average fewer loci have the same allele in all lines (i.e. hidden loci), which increases the overall number of detected loci and power of the test. (B) The mean p-value for the selective scenario Q_1 decreases with the number of lines. Simulation parameters: $\sigma = Nsg = 1$ with mean trait contribution $g = 0.1$ per locus, $k = 20$. Parameters of the gamma distribution of trait contributions are $\alpha = 2$, $\beta = 20$.

The simulations show that the log-odds score for the selective model Q_1 increases with the number of lines while the mean p-value decreases (see Figure 1). The increase in score is largest from 2 to 3 lines, where the increase in score of 50% corresponds to an average increase of 50% in the number of observed loci. The dependence of the score S on the number of lines k is approximately given by $S(k) = 2S_2(1 - 2^{-k+1})$, where S_2 is the score for $k = 2$ lines. Thus the score for large k is twice the score for two lines.

In the rest of the paper we separate the effect of the increased number of observed loci from other qualitative differences between two and more than two lines by fixing the number of *diverged* loci, which we call k_{div}.

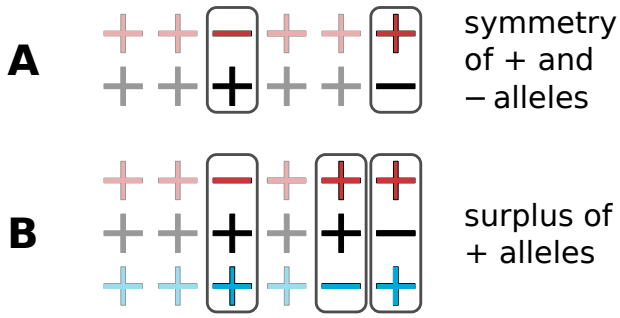

FIG. 2. **Comparing allele statistics in two and three lines** (A) In our example of two lines evolving under identical selection strength s, the value of s cannot be inferred, since for the diverged loci + and −-alleles appear with the same probability in the two lines. (B) On the other hand, for three lines under the same selection strength (for a high trait value), a surplus of +-alleles is seen, which allows to deduce the absolute selection strength in all three lines. The loci with faded color are not observed in experiment.

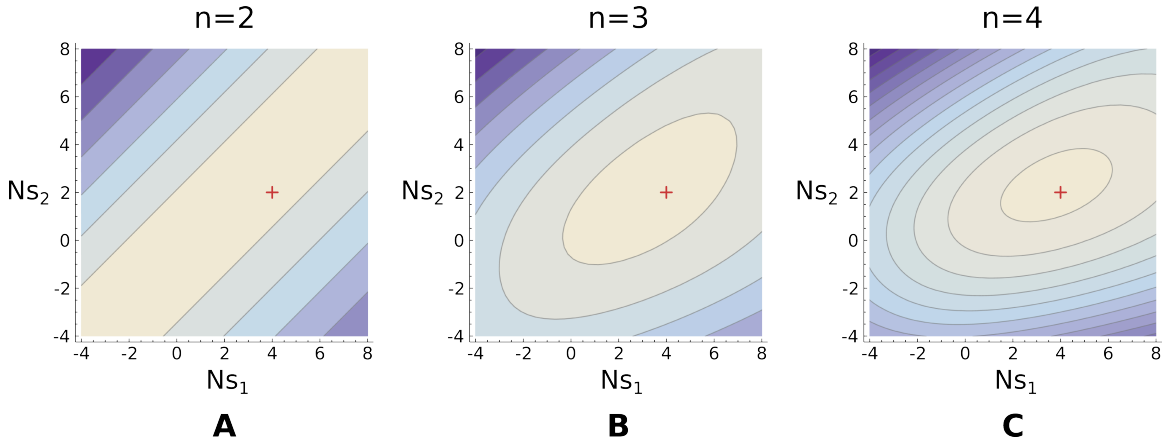

FIG. 3. **Inferring selection strength at equilibrium.** The log-likelihood given by (2) at large evolutionary times (equilibrium) for the two selection strengths Ns_1 and Ns_2 shows qualitatively different behavior for $n = 2$ lines and $n > 2$ lines. (A) For $n = 2$, scenarios with the same $\Delta s = s_1 - s_2$ have the same likelihood, making it impossible to determine the absolute values of Ns_1 and Ns_2. (B) For $n = 3$ lines, however, the likelihood shows a unique peak around the true values of the selection parameters (red cross at $Ns_1 = 4, Ns_2 = 2$). (C) Increasing the number of lines to $n = 4$ narrows this peak in the likelihood. The contour plot is scaled such that there is an additive difference in the log-likelihood of 2 between two adjacent contour lines. Simulation parameters: $k_{\mathrm{div}} = 20$ diverged loci; trait contributions g_i drawn from a gamma distribution with parameters $\alpha = 2$ and $\beta = 20$.

V. INCREASED SCOPE IN MORE THAN TWO LINES

Besides the increase in statistical power due to more observed loci, there is another important difference between the equilibrium statistics of two lines and the statistics of more than two lines. The statistics of two lines depends only on the *difference* between selection strengths of the two lines, whereas the statistics of more than two lines depends on the selection strengths on the trait in all lines (see Tables I and II). As a result, the inference of absolute rather than relative selection strength in equilibrium requires more than two lines. We first consider the following special case as an example. Consider two lines with biallelic loci contributing to a trait. The trait evolves under the same positive selection strength $s_1 = s_2$ in the two lines. Looking at loci with alleles differing by state, the two configurations $(+-)$ and $(-+)$ occur with the same frequency, (see Figure 2). This is the same result that would be obtained under neutral evolution. There are also many loci with a $(++)$ configuration, but these are not experimentally detectable

from two-line crosses (Figure 2). So with a two-line test only the difference in selection strength between the two lines can be determined. The general case is described in Table I, where the probabilities of the diverged configurations of a locus are always a function of the difference $\Delta s = s_1 - s_2$. However, for three lines under the same selection strength $s_1 = s_2 = s_3$, one would mostly observe loci where two of the lines have the +-allele, but rarely a configurations with two −-alleles. So this allele distribution differs from the neutral distribution even in the absence of relative differences of selection strength between the lines. The configurations $(++-)$, $(+-+)$ and $(-++)$ would appear with the same frequency, pointing towards equal selection strength in all lines. An analogous argument applies to general selection strengths: In Table II not only the differences, but also the sums of the selection strengths appear in three-line statistics, allowing to infer absolute selection strengths instead of relative differences. The same argument applies for any number of lines greater than two.

q_1	q_2	relative probability
+	-	$e^{N(s_1-s_2)g}$
-	+	$e^{-N(s_1-s_2)g}$

TABLE I. **Relative probabilities for the two possible configurations of a diverged locus for two lines.** In both cases only the difference of selection strength $s_1 - s_2$ enters the probability (2) for a given configuration.

q_1	q_2	q_3	relative probability
-	-	+	$e^{N(-s_1-s_2+s_3)g}$
+	-	+	$e^{N(+s_1-s_2+s_3)g}$
-	+	+	$e^{N(-s_1+s_2+s_3)g}$
+	-	-	$e^{N(+s_1-s_2-s_3)g}$
-	+	-	$e^{N(-s_1+s_2-s_3)g}$
+	+	-	$e^{N(+s_1+s_2-s_3)g}$

TABLE II. **Relative probabilities for the six possible configurations of a diverged locus for three lines.** Configurations that show no difference between two lines (e.g. $(q_1, q_2) = (+, +)$) can now be observed (as in $(q_1, q_2, q_3) = (+, +, -)$), such that not only the differences but also sums of the selection strengths appear in the allele probabilities.

Inference of absolute selection strengths in equilibrium: This simple result has a drastic consequence for the inference of selection strength and testing for selection. When inferring selection strengths in two lines, the likelihood of s_1, s_2 depends only the relative selection strength $\Delta s \equiv s_1 - s_2$. To illustrate this point, we look at the likelihood of the selection parameters s_i under the model (2), varying the number of lines n. For this, we draw allele configurations $(q_1, \ldots, q_n)$ for $k_{\text{div}} = 10$ diverged loci from distribution (2), setting $Ns_1 = 4$ and all other selection strengths $Ns_2 = \ldots = Ns_n = 2$. Given this set of observed loci, we plot the log-likelihood $\sum_{i=1}^{k_{\text{div}}} \log(P(q_{1,i}, \ldots, q_{n,i}|g_i))$ as a function of the two selection parameters Ns_1 and Ns_2 for $n = 2$, 3 and 4 lines (see Figure 3). For $n = 2$ the likelihood is constant for lines of constant Δs, making it impossible to determine the absolute values of s_1 and s_2. For $n = 3$, however, the likelihood is peaked approximately around the correct values $Ns_1 = 4$, $Ns_2 = 2$. This peak becomes even narrower when 4 lines are considered, allowing to determine the selection parameters more precisely.

Numerical simulations: To test the ability of the test to distinguish different neutral and selective scenarios using multiple lines, we perform numerical simulations on artificial data. We set the number of lines to $n = 3$ and fix the number of diverged loci to $k_{\text{div}} = 10$. To the above defined completely neutral scenario P_0 and the lineage-specific selection scenario Q_1 we add the scenario P_s, where we allow for a uniform non-zero selection strength in all lines ($s_1 = s_2 = s_3 = s$) and the scenario Q_3, where all three lines are under selection but in different directions ($s_1 = +s$, $s_2 = -s$, $s_3 = -s$). We also test the only possible selective two-line scenario Q_{two} with a relative difference Δs of selection strength between the lines against the neutral scenario P_{two} with $\Delta s = 0$. Analogously to the simulations performed in the previous section, scenarios Q_1 and P_s, Q_1 and P_0, Q_3 and Q_1, and Q_{two} and P_{two} are compared against each other. For the test of scenarios Q_1 against P_s and the scenarios Q_3 against Q_1 in each case the selection

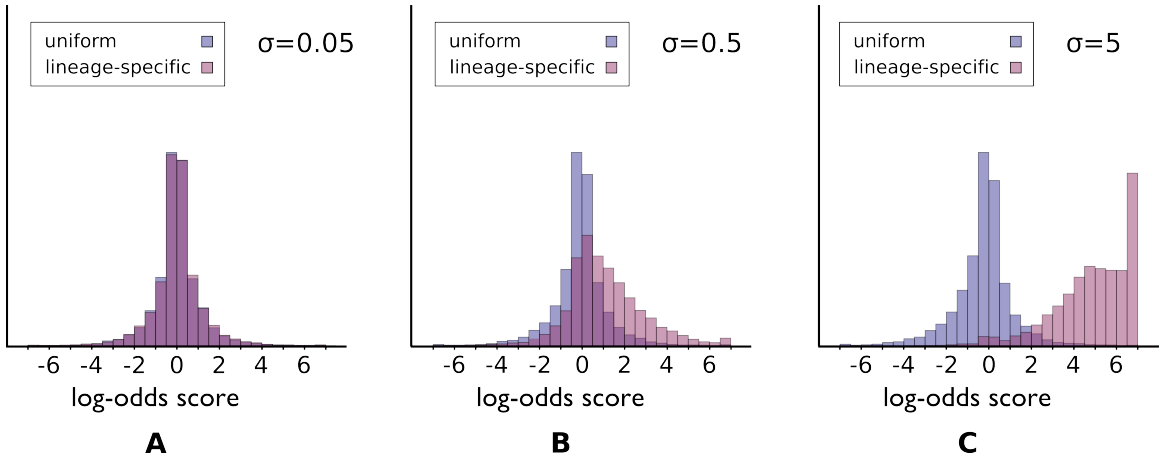

FIG. 4. **Score statistics for a trait in different evolutionary scenarios.** The figure shows the distribution of the log-odds score (3) for many realizations of the allele statistics arising under the uniform selection scenario P_s and the lineage-specific selection scenario Q_1 for different values of the selection coefficient $\sigma = Nsg$. (A) For small selection coefficients ($\sigma = 0.05$), the score distribution (3) under both scenarios is nearly identical. In this case, the two evolutionary scenarios are hard to distinguish. (B and C) As the selection strength s increases, the score distributions clearly separate. Simulation parameters: three lines, $\sigma = Nsg = 0.05$, 0.5 and 5 with mean trait contribution $g = 0.1$ per locus, $k_{\rm div} = 10$. Parameters of the gamma distribution of trait contributions are $\alpha = 2$, $\beta = 20$.

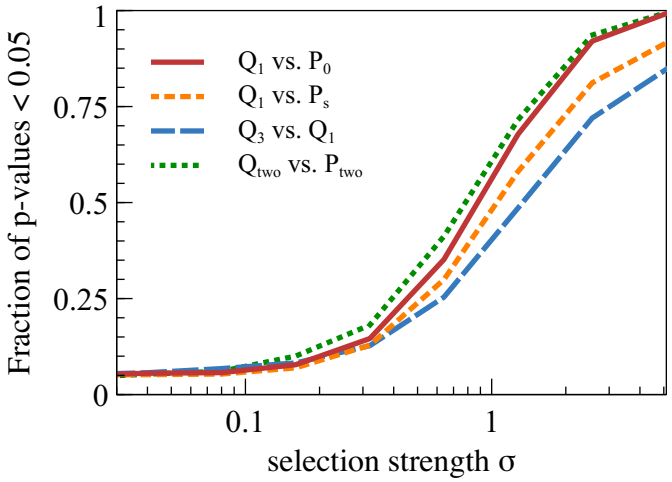

FIG. 5. **Statistical significance of the tests at different levels of selection strength.** For three lines, the allele statistics generated under different evolutionary scenarios (Q_1, Q_3) are probed with tests for different hypotheses. We test selective scenarios against the neutral hypothesis (Q_1 vs P_0), as well as different selective scenarios (Q_1 tested against uniform selection P_s, and the lineage-specific scenarios Q_1 and Q_3 compared against each other). The fraction of realizations where the log-odds score is statistically significant ($p < 0.05$, see text) rises steeply with increasing selection strength (selection coefficient $\sigma = Nsg$ per locus). For two lines, the test for relative difference in selection strength ($Q_{\rm two}$ vs $P_{\rm two}$) shows a similar statistical power (green dotted line). Simulation parameters as in Figure 4.

strength for both scenarios is chosen to yield the same mean phenotype $\bar{G} = (G_1 + G_2 + G_3)/3$.

The simulations show that the log-odds score can clearly distinguish selective and neutral scenarios (see Figure 4 and 5), uniform selection and lineage-specific selection, as well as between different lineage-specific selection scenarios. As expected, the sensitivity of the test increases with selection strength. The test works in a reasonable parameter range, allowing to infer selection strength with only few loci available ($k \gtrsim 4$ loci for $Nsg = 1$) and a reasonable selection strength ($Nsg = 1$ corresponds to a probability of 0.88 for a locus to be in the + state).

Short time dynamics: We also consider the allele statistics in the limit of short evolutionary times since the last common ancestor. In this limit each locus has changed state at most once. The detailed calculations are given

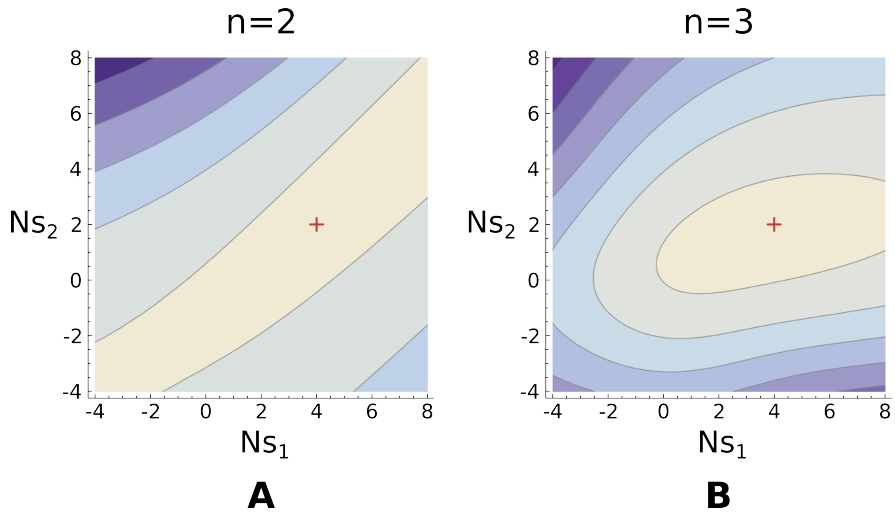

FIG. 6. **The log-likelihood for selection strengths at short evolutionary distances.** (A) For $n = 2$, the log-likelihood of the selection parameters has a very flat maximum, resembling the case of long evolutionary times in Figure 3. Knowledge of the ancestral allele is assumed. Even with this information available, it is hardly possible to estimate the selection parameters due to the flat maximum. (B) For $n = 3$ lines, however, the estimate of the selection parameters is again possible, even when, as in this case, no further knowledge of the ancestral alleles of the loci is assumed. Simulation parameters: $Ns_1 = 4$ and $Ns_2 = 2$ for $n = 2$ and $t_1 = t_2$, $Ns_1 = 4$ and $Ns_2 = Ns_3 = Ns_{23} = 2$ for $n = 3$. Other simulation parameters are as in Figure 3.

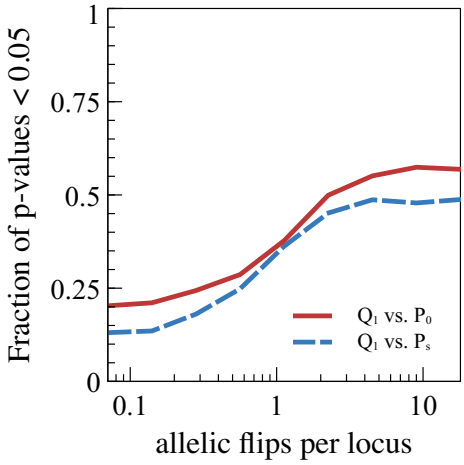

FIG. 7. **The power of the equilibrium test at short times.** The significance of the three-line equilibrium selection test decreases somewhat with decreasing number of allelic flips per locus (corresponding to shorter evolutionary timescales). However, both at intermediate times and even for short evolutionary times the equilibrium test retains some of its power. Simulation parameters: evolutionary scenario Q_1 tested against P_0 and P_s, $\sigma = Nsg = 1$, number of diverged loci $k_{\mathrm{div}} = 10$.

in the appendix. Since the ancestral states of the loci and the phylogenetic tree of the lines has to be considered, the general results for n lines are unwieldy. Here, we compare the cases of $n = 2$ and $n = 3$.

For two lines the ancestral state c cannot be inferred, since both final configurations $(q_1, q_2) = (+-)$ and $(-+)$ can be reached from either ancestor $c = \pm$. But even if the ancestral states were known, the allele distribution for two lines suffers from similar problems as in the case of long evolutionary times, see Figure 6. Given the ancestral states, the probability for a mutation in line a is $P(a|g, c) = \frac{s_c(g,Ns_a)}{s_c(g,Ns_1)+s_c(g,Ns_2)}$, where $s_c(g, Ns) = \frac{-2Nscg}{1-e^{2Nscg}}$ is the Kimura transition probability [38] and Ns_a is the selection strength of the trait in the line undergoing a mutation. The resulting allele statistics depends on absolute selection strengths s_1, s_2 (rather than just the difference $\Delta s = s_1 - s_2$). Yet, for $\Delta s = 0$ the allele distribution equals that of the neutral case $s_1 = s_2 = 0$. Also simulations of the likelihood landscape for $\Delta s \neq 0$ give a very flat likelihood peak, resembling the equilibrium case and making estimates of the

absolute selection parameters very unreliable (Figure 6). This suggests that also for short evolutionary times inference of absolute selection strengths on a trait based on two lines is hard, that is, it requires a large number of loci. For unknown ancestral alleles one can average over the two possible ancestral alleles (see eq. (10) in the appendix).

Again the situation changes when considering three or more lines. For three lines, four of the six possible configurations can be assigned a unique ancestral allele ($c = -$ for $(q_1, q_2, q_3) = (+--)$ and $(-+-)$, $c = +$ for $(-++)$ and $(+-+)$ with a phylogenetic tree as in Figure 9). Here, we can write, without knowledge of ancestral states, the allele distribution under the most general selection scenario as

$$Q_s(q_1, q_2, q_3|g) = \frac{1}{Z}\Big(t_2 P(l) s_l(g, Ns_{q=-l}) + \delta_{q_1,q_2} t_1 \cdot [P(l)s_l(g, Ns_3) + P(-l)s_{-l}(g, Ns_{12})]\Big), \quad (4)$$

where t_1 and t_2 denote the branch lengths of the phylogenetic tree (see Figure 9 in Appendix I), $l = q_1 + q_2 + q_3 = \pm 1$, $P(c)$ is the prior probability of an ancestral state $c = \pm 1$, $Ns_{q=-l}$ denotes the selection strength of the line with the minority allele (e.g. Ns_3 for the configuration $(--+)$) and $Z = \sum_{q_1', q_2', q_3' = \pm 1} Q_s(q_1', q_2', q_3'|g)$. Again, the two states with $q_1 = q_2 = q_3$ are excluded. Figure 6 shows a plot of the log-likelihood for three lines with a well-defined maximum, indicating that it is possible to infer absolute selection parameters from three-line data. Unlike the case of two lines, this was done without additional information on the ancestral allelic states.

Time to equilibration: The assumption that allelic states follow their equilibrium statistics reached after a long evolutionary time need not hold in a particular application. The equilibration time of the QTL statistics depends, besides the mutation rate, on two factors: the strength of selection and the size of mutational targets. If a QTL locus is an entire gene with its flanking regulatory regions, there might be many mutations at different nucleotides which affect the locus in similar ways (e.g. increase the expression level of a particular gene) [44]. But the size of this mutational target is hard to quantify since usually not much is known about the number of possible mutations affecting a QTL locus. One particularly interesting example of large mutational targets are plants, where a large part of the genome consists of transposable elements [45, 46]. These elements can insert themselves into a gene or its regulatory flanking regions and alter gene expression [47, 48] or even lead to the deletion of whole genes [49].

Here, we probe the statistical power of the equilibrium test over different evolutionary times. The log-odds score appropriate in the limit of short evolutionary times is derived in the appendix. The relevant quantity to characterize the evolutionary time since the last common ancestor of two lines in the neutral case is $t\mu$, where t is the evolutionary time in generations and μ the mutation rate per locus per generation. $t\mu$ quantifies the average number of allelic changes fixed in the population ($+ \to -$, $- \to +$) per locus. If $t\mu \gg 1$, each locus has changed state many times since the divergence of two lines (regime of long evolutionary times), while for $t\mu \ll 1$ most loci did not change state and the diverged loci mostly underwent a single mutation only (regime of short evolutionary times). Multiple changes of a locus can be realized by changing the same locus in different lines. This possibility is supported by experimental evidence in maize, where in many cases the same loci have undergone evolutionary changes in many different lines [31].

We simulate the regime of intermediate evolutionary times by altering $t\mu$ at a constant selection strength $Ns = 10$ and a constant number of diverged loci $k_{\mathrm{div}} = 10$ (which is smaller than the total number of QTL loci k for $t\mu \ll 1$) for the Q_1-scenario tested against the P_0 and P_s scenario (see Figure 7). The statistical power decreases gradually in both scenarios when going from long to very short evolutionary times. Yet, the test retains some of its statistical power even as μt goes to zero, i.e. each diverged locus changed state only once. For the allele statistics and the resulting log-odds score at short evolutionary times, see appendix I.

VI. MULTIPLE TESTING CORRECTIONS

When testing for selection on a quantitative trait, one particular subtlety needs to be considered. As emphasized by Orr [27], a large difference of a trait between two lines is not sufficient evidence for lineage-specific selection. Often, traits in QTL experiments are picked from a larger pool of traits; among those, traits that diverged markedly between lines are chosen for further analysis, since this difference hints at lineage-specific selection. However, in a sufficiently large set of traits, neutral evolution alone would produce traits differing between the lines. In such a trait, one would also observe an imbalance of alleles enhancing the trait value in one line and reducing it in the other. The bias in trait difference and allele statistics resulting from a non-random choice from a set of traits is called ascertainment bias. Ascertainment bias can lead to non-neutral evolution being attributed to a trait that evolved neutrally along with a set of other neutrally evolving traits.

There are two ways to correct for this effect. If the total number of traits from which the observed trait is taken is known explicitly, we are faced with a standard multiple-testing problem. We consider this case first. However, if the

trait is chosen from an ill-characterized set of traits, the situation is different. We follow the approach of Orr [27] and consider the allele statistics conditioned on the observed phenotypic difference. We will see that in this case, again, there is a drastic difference between two and more than two lines.

Holm-Bonferroni correction: If the total number of observed traits is known, a standard multiple-testing correction can be applied. An example is gene expression levels, where traits are analyzed on a genome-wide level and the number of genes is known [29]. A suitable multiple-testing correction for this case is the Holm-Bonferroni correction, which has the advantage that no independence of the different hypotheses tested needs to be assumed. This is particularly important in QTL analysis, since different traits can be affected by the same genetic loci (an example will be discussed in the next section). The Holm-Bonferroni correction controls the familywise error rate (FWER), i.e. it controls the false positive rate not only for a single trait but a whole set of traits. If there are m traits for which scenario Q is tested against the null hypothesis of scenario P, one calculates the log-odds score $S_{Q,P}$ (3) and the corresponding p-values p_i for all m traits. The traits are then ranked according to their p-values with the highest p-values first. Now one searches for the first trait i for which $p_i > \alpha/(m+1-i)$, where α is the significance threshold for the familywise error rate. Scenario P can then be rejected for the traits $1,\ldots,i-1$, but not for the traits $i,\ldots,m$.

Conditioning on the phenotypic difference: Often, the size of the pool from which traits are picked is not known. Most traits from this pool remained unnoticed simply because they showed little difference between the lines and were not identified as possible traits for investigation. The proposal of Orr [27] for this case is to condition the allele distribution on the observed phenotypic difference, that is to restrict all models to the allele configurations with the phenotypic difference observed between two lines. In doing so, the part of the evidence for selection that comes from the phenotypic difference between two lines is discarded. Orr writes the phenotypic difference as $R = \sum_{i=1}^{k} g_i(q_{1,i} - q_{2,i})$ for the case of two lines. We generalize this notion to the case of n lines and denote the maximal phenotypic difference between two lines $R_{max} = \sum_{i=1}^{k} g_i(q_{1,i} - q_{2,i})$, where we order the lines such that line 1 has the largest phenotypic value $G_1 = \sum_{i=1}^{k} g_i q_{1,i}$ and line 2 has the lowest phenotypic value G_2.

Our next step is to calculate the allele statistics conditioned on a particular value of $R_{\max}$. This statistics can then be used in place of the allele statistics $P(q_1,\ldots,q_n|g)$ under neutrality in the log-odds score (3). Our calculation is based on the principle of maximum entropy. This general principle applies to situations with incomplete knowledge about the probability distribution $p(x)$ of some variable x. This distribution must be consistent with any prior information on x one might have (for instance the mean value of x), but otherwise it should be as unbiased as possible. The principle of maximum entropy posits that the distribution which best describes the state of our knowledge is the distribution which maximizes the information entropy $-\sum_x p(x)\ln p(x)$ with respect to $p(x)$, subject to the constraints resulting from prior information. Stated in this form first by E. T. Jaynes [50], the principle of maximum entropy already appears at the core of statistical physics, where the distribution over configurations x of a physical system are constrained the mean energy $\langle E(x)\rangle = \sum_x E(x)p(x)$. The maximum entropy distribution in this case turns out to be the Boltzmann distribution $p(x) \propto e^{-\beta E(x)}$, where β is determined by the mean value of the energy $E(x)$. Other applications of the principle of maximum entropy are in image reconstruction [51] and language modelling [52]. In the context of quantitative traits, the principle of maximum entropy and the associated calculus of Boltzmann distributions has been used to estimate unobserved allele frequencies and to infer selection from trait observables [32–35, 40, 53–58]. Here, we use the principle of maximum entropy to derive the allele distribution conditioned on the largest phenotypic difference $R_{\max}$. A pedagogical example is given in Appendix II.

We consider a distribution $P_{s,h}(q_1,\ldots,q_n)$ with an additional parameter h determining the value of $R_{\max}$. This distribution is obtained by maximizing the information entropy

$$\begin{aligned} H(P) = &- \sum_{q_i=\pm 1} P_{s,h}(q_1,\ldots,q_n)\log\frac{P_{s,h}(q_1,\ldots,q_n)}{P_s(q_1,\ldots,q_n)} \\ &+ \lambda_0\left(\sum_{q_i=\pm 1} P_{s,h}(q_1,\ldots,q_n) - 1\right) \\ &+ h\left(g\sum_{q_i=\pm 1}(q_1-q_2)P_{s,h}(q_1,\ldots,q_n) - R_{\max}\right) \end{aligned} \tag{5}$$

with respect to $P_{s,h}(q_1,\ldots,q_n)$. The sum over all possible states $q_i = \pm 1$, $i = 1,\ldots n$ for a given locus again excludes the two unobserved states with $q_1 = \ldots = q_n = \pm 1$. We have two Lagrange multipliers λ_0 and h to implement the normalization constraint and the constraint on the largest phenotypic difference $R_{\max} = \sum_i g_i(q_{1,i} - q_{2,i})$. Setting

the derivative of the information entropy (5) with respect to $P_{s,h}(q_1,\ldots,q_n)$ equal to zero gives

$$P_{s,h}(q_1,\ldots,q_n) = \frac{e^{hg(q_1-q_2)+gNs\sum_{i=1}^n q_i}}{\sum_{\{q_1',q_2',\ldots,q_n'=\pm 1\}} e^{hg(q_1-q_2)+gNs\sum_{i=1}^n q_i'}}\,. \tag{6}$$

The parameter h is determined by the phenotypic difference $R_{\max}$: in line 1 alleles have a higher probability to be in the +-state while in line 2 alleles have a higher probability to be in the −-state.

This conditioned allele distribution can now used in place of the two scenarios P_0 and P_s to account for ascertainment bias. The resulting log-odds score

$$S_{Q,P} = \sum_{i=1}^{k} \ln\left(\frac{Q(q_{1,i},q_{2,i},\ldots,q_{n,i}|g_i)}{P_{s,h}(q_{1,i},q_{2,i},\ldots,q_{n,i}|g_i)}\right) \tag{7}$$

depends the parameter h; extremizing the score with respect to h sets the expected value of the phenotypic difference under the conditioned model $P_{s,h}$ equal to the phenotypic difference observed in the data.

In the case of $n = 2$ lines it turns out that the probabilities for the two observable allelic states under conditioning, $P_{s,h}(q_1,q_2) = e^{hg(q_1-q_2)}/C$ (where the uniform selection parameter s drops out), are the same as for the selective model at equilibrium, $Q(q_1,q_2) = e^{\Delta sg(q_1-q_2)}/C$, (see also Table I). Setting h such that the expected phenotypic differences are the same under the conditioned neutral model and the selective model at equilibrium, the statistics of under ascertainment and under selection are exactly the same[59]. As a result, the log-odds score comparing evolution under selection at equilibrium with the neutral statistics conditioned on the observed phenotype is exactly zero. For two lines at equilibrium it is not possible to statistically distinguish neutral evolution under conditioning on the phenotypic difference on a trait from the effect of selection.

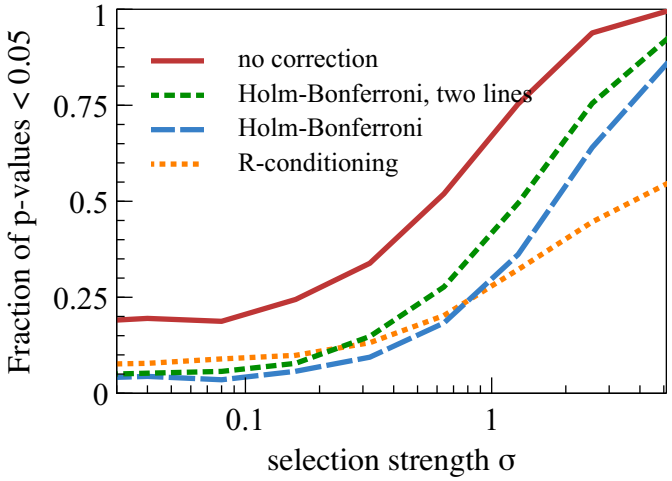

FIG. 8. **Comparing the statistical significance of tests with different multiple-testing corrections.** For three lines, the corrected tests both have a lower statistical significance but they can control the false positive rate under ascertainment bias when testing many different traits. The high false positive rate of 0.18 without correction is reduced to 0.04 for the Holm-Bonferroni correction and to 0.076 for the conditioning on $R_{\max}$ (which can be read of as the fraction significant outcomes under neutrality for $\sigma \to 0$ at the very left of the plot). For high selection strength the Holm-Bonferroni correction has a higher true positive rate than the conditioned test. For two lines, using the Holm-Bonferroni correction allows to distinguish scenarios Q_{two} and P_{two}, contrasting the case under conditioning $R_{\max}$ where no distinction is possible at all. Simulation parameters: $m = 5$ traits, significance threshold $\alpha = 0.05$ in all three cases. The other simulation parameters are as in Figure 4.

Again, the situation is fundamentally different when considering more than two lines. The allele statistics in the selective scenario in equilibrium for more than two lines differs from the neutral scenario, even when the ascertainment bias is accounted using the maximum-entropy result (6) (as in the inference of absolute selection strength in section II). For more than two lines, the score (7) gives non-zero results both at equilibrium and at short evolutionary times. There is only one exception, namely the particular selection scenario ($s_1 = +s, s_2 = -s, s_3 = 0, \ldots, s_n = 0$) that is not distinguishable from neutral evolution conditioned on R_{max}.

To test this approach to the multiple-testing problem and compare it to the Holm-Bonferroni correction, we simulate a multiple-testing scenario where a trait is picked from a larger set of traits. First, we draw alleles for $m = 5$ traits for three lines all evolving neutrally under scenario P_0. Then the traits are sorted according to the maximal phenotypic difference $R_{\max}$ between any two of the three lines. The trait with the highest $R_{\max}$ is tested using the selective scenario Q_1 against the neutral scenario. We do this in three ways: using the score (3) without a multiple-testing correction, by applying the Holm-Bonferroni correction assuming the number of traits is known, and by conditioning on $R_{\max}$ using (7). Repeating this procedure many times over, we compute the false positive rate for all three approaches. This corresponds to the leftmost points of Figure 8. Second, we generate the allele statistics of one trait under the selective scenario Q_1 and for the other traits under the neutral scenario P_0. Then, we determine how often the trait under selection is correctly identified by the different approaches with a p-value less than 0.05 ($0.05/m$ for Holm-Bonferroni). Figure 8 shows that, as expected, a test without correction yields the highest rate of true positives. Yet, it suffers from a high false positive rate of 0.18, since many of the neutrally evolving traits which happen to have a high $R_{\max}$ are identified as being under lineage-specific selection. The Holm-Bonferroni method has the second highest significance rate for high selection coefficients σ and the lowest false positive rate of 0.040 (which is lower than $\alpha = 0.05$ since α controls the familywise error rate of all five traits). The conditioning on $R_{\max}$ has the lowest true positive rate for high values of the selection coefficient σ and a false positive rate of 0.076 which is clearly lower than in the uncorrected case but still elevated compared to the Holm-Bonferroni correction. The true positive rate under conditioning is higher than for Holm-Bonferroni for small σ only because of the elevated false positive rate (the family-wise error and the significance threshold is not directly comparable in the two cases). Thus in those cases where the size of the pool of traits is known or can be estimated, the Holm-Bonferroni correction is to be preferred over conditioning on the phenotypic difference (as expected since this condioning discounts the evidence from the phenotypic differences). For two lines, simulations with the Holm-Bonferroni correction show that, unlike under the conditioning on $R_{\max}$, scenarios Q_{two} and P_{two} are distinguishable (see Figure 8).

While the maximum phenotypic difference $R_{\max}$ is a plausible observable on the basis of which traits can be selected from a larger pool, it is by no means the only one. For instance, with three lines, traits could be selected based on the difference between trait in line 1 and the trait mean in line 2 and 3, $R_\Delta = G_1 - \frac{G_2+G_3}{2}$. One would use this observable when looking specifically for traits with lineage-specific selection acting on line 1. For $s_2 = s_3 \equiv s_0$ the fitness can be written

$$F(G_1, G_2, G_3) = \bar{s}(G_1 + G_2 + G_3) + \hat{s}\left(G_1 - \frac{G_2 + G_\gamma}{2}\right) \tag{8}$$

with $\bar{s} = (s_1 + 2s_0)/3$ and $\hat{s} = s_1 - \bar{s} = 2(s_1 - s_0)/3$. The maximum entropy distribution conditioned on R_Δ is up to a normalizing constant $\exp\{\lambda R_\Delta\}$ and thus differs from the equilibrium distribution $\propto \exp\{\beta F(G_1, G_2, G_3)\}$, except in the special case $\bar{s} = 0$. This illustrates again that we can infer the absolute strength of selection from three lines, but not from two.

VII. SELECTION ON PLANT QUANTITATIVE TRAITS

We apply the multiple-line selection test to two studies of plant quantitative traits. Our first example is based on QTL data on corolla (petal) sizes in three different plant species of the genus *Mimulus*. *M. guttatus*, *M. platycalyx*, *M. micranthus* are labelled lines 1, 2 and 3 respectively. At each locus detected in [25], it turns out there are two alleles with very similar trait contributions (within experimental error), and one allele with a significantly different contribution. If there is a single high allele, we assign this allele the +-label, while the two other alleles are assigned the label −, and vice versa. The resulting values for the corolla width and corolla length trait are in Table III. We calculate the log-odds score for the pairwise comparisons of the evolutionary scenarios introduced above. Where applicable, we use the Bayesian information criterion described above to correct the scores for different numbers of free parameters. (This leaves the p-values unaffected.) Testing against neutrality (scenario P_0) and against uniform selection strength (scenario P_s) we use the test either with the conditioning on R or with the Holm-Bonferroni correction. In the first case, we condition the null model on the pair of lines with the highest phenotypic difference for each trait. For the Holm-Bonferroni test we take $m = 5$ as the total number of traits for this dataset since there are 5 different traits incorporated in the QTL experiment in [25]. Since the trait pool size may well be higher than this, we test the robustness of our results by also identifying the number of hidden traits m_h required to make the result statistically insignificant.

We start with the corolla width trait, where 7 QTL have been identified along with their trait contributions [25]. First we apply the conditioned allele distribution according to eq. (6). For this trait we condition the null models P_0 and P_s on R_{12}. First we compare the scenario Q_3 (line 1 under selection for a large trait value, line 2 and 3 for a small trait value) against P_0 (neutral evolution under ascertainment) which gives a log-odds score (3) of $S_{Q_3,P_0} = 1.05$

trait contribution g_i	*M. gutt.*	*M. platy.*	*M. micr.*
corolla width [mm]:			
0.41	-	+	-
0.74	+	-	-
0.39	+	-	+
0.59	+	-	-
0.28	+	+	-
0.64	+	-	-
0.37	+	-	+
corolla length [mm]:			
0.67	+	-	-
0.41	+	+	-
0.21	+	-	+
0.60	+	-	-
0.27	-	+	+
0.51	+	+	-

TABLE III. QTL trait contributions to two flower traits of the Mimulus species *M. guttatus*, *M. platycalyx* and *M. micranthus* estimated from [25].

in favour of the selective scenario. We test the significance of this score by repeated simulations under the scenario P_0 with the estimated selection parameter h on the given set of observed trait contribution g_i (see Table III). For each configuration drawn from P_0 we sort the lines according to their phenotypes G_i. In this way we account for the possibility that under neutrality fluctuations create patterns of lineage-specific selection in any of the lines (rather than only in what we called line 1 here). We obtain a p-value $p = 0.034$, which favours the selective model Q_3 over the neutral null hypothesis. Next, we test whether lineage-specific selection is required to explain the observed trait differences by comparing the Q_3 scenario against the full null model P_s. We obtain a score of $S_{Q_3,P_s} = 1.51$ ($p = 0.031$), which points towards lineage-specific selection. We note that the difference between the two scores S_{Q_3,P_0} and S_{Q_3,P_s} is exactly the score between the two null models S_{P_s,P_0}.

The unconditioned test together with the Holm-Bonferroni correction yields very similar results. Testing scenario Q_3 against P_0 we obtain a score of $S_{Q_3,P_0} = 3.55$ and a corresponding p-value of $p = 0.014$. Since this is the smallest p-value of all traits for the comparison of these scenarios, we apply a more stringent p-value cutoff $p < \alpha/m = 0.01$ for $\alpha = 0.05$ and $m = 5$. The result is just above the chosen significance threshold, but the cutoff for the Holm-Bonferroni test is stricter than the cutoff for the conditioned test for the same α (see numerical simulations section). For scenario Q_3 and P_s the score $S_{Q_3,P_s} = 3.03$ ($p = 0.018$) again leads to a similar result as the conditioned test.

These results are in agreement with the different reproductive modes of these species [25]: line 1 reproduces predominantly by outcrossing (so that large floral characters are needed to attract pollinators), whereas line 2 and line 3 are self-pollinating. In the latter species, large petals give no advantage for reproduction, but nevertheless require resources to develop and maintain.

Of course one can also test different lineage specific scenarios against each other. For this particular trait, testing scenario Q_3 against scenario Q_1 yields a score $S_{Q_3,Q_1} = 1.42$ ($p = 0.035$) in favour of scenario Q_3 consistent with the previous results. We also apply the general three-line scenario Q_{Full}, where selection parameters Ns_1, Ns_2 and Ns_3 are determined independently for each line. Computing the set of selection parameters by maximum likelihood, we obtain $Ns_1 = 1.04$, $Ns_2 = -1.26$ and $Ns_3 = -1.36$. These are close to the result of the Q_3-scenario ($Ns = 1.23$), supporting this restricted scenario. However, the score for the comparison between the general and the neutral scenario P_0 is not statistically significant ($S_{Q_{\text{Full}},P_0} = -0.88$, $p = 0.33$ under conditioning and $S_{Q_{\text{Full}},P_0} = 0.64$, $p = 0.14$ with Holm-Bonferroni). The large number of degrees of freedom of the general scenario reduces the statistical power of the resulting test.

Next, we consider the corolla length trait, where 6 QTL were observed (see Table III). Again we apply the conditioned allele distribution first. This time the maximal phenotypic difference is R_{13}. Here, the comparison to the neutral null-model yields the score $S_{Q_3,P_0} = -0.53$ ($p = 0.30$), slightly but not significantly in favour of the neutral hypothesis. Also the test for lineage-specific selection yields no significant result ($S_{Q_3,P_s} = -0.54$, $p = 0.36$). The Holm-Bonferroni procedure produces a significant result tested against the uniform selection scenario ($S_{Q_3,P_0} = 1.84$, $p = 0.097$ and $S_{Q_3,P_s} = 1.84$, $p = 0.048$). Similarly, the test against the other lineage-specific scenario Q_1 gives no significant result ($S_{Q_3,Q_1} = -0.16$, $p = 0.19$) as well as the general, unrestricted three-line scenario when tested against the neutral scenario ($S_{Q_{\text{Full}},P_0} = -1.68$, $p = 0.14$ in favour of the neutral model under conditioning and

$S_{Q_{\rm Full},P_0} = -0.21$, $p = 0.24$ for Holm-Bonferroni). A summary of the results can be found in Table V. We note that several of the QTL controlling corolla width and length overlap with each other (see [25]).

For comparison, we also apply Orr's sign test [27] (not the equal effects version) to this dataset. Since the Orr-Test is a two-line test, we apply it to the two lines with the largest phenotypic difference, where one would expect the strongest signal for selection. Following Orr, the trait contributions g_i are taken from a gamma distribution whose parameters for each trait are estimated by maximum likelihood. Then the probability to find at least the observed number of +-alleles in the high line given the observed phenotypic difference R or greater is calculated according to eq. (4) in Orr's paper [27]. For the corolla width trait 5 out of 6 diverged loci in line 1 and 2 have the +-allele. Here, the Orr-test shows no significant result ($p = 0.42$). Also for the comparison of line 1 and line 3 the test yields no significant result (p=0.29). For the corolla length trait we have 4 out of 5 diverged loci in the + direction between line 1 and line 3 and 3 out of 4 diverged loci in the + direction between lines 1 and 2. Again the Orr-test gives no significant result in either case ($p = 0.48$ and $p = 0.72$, respectively). In order to apply the more recent selection test of Rice and Townsend [28], additional data from a mutation accumulation experiment for the observed Mimulus traits would be necessary.

trait contribution g_i	B73	B97	CML254	Ki14
GDDTA [GDD]:				
4.73	-	-	+	+
3.85	-	-	+	-
4.43	-	-	+	-
11.13	-	-	+	+
GDDTS [GDD]:				
6.33	-	-	+	+
6.20	-	-	+	-
4.68	+	-	+	+
5.68	-	-	+	+
plant heightbp [cm]:				
1.10	-	-	+	+
1.25	+	+	-	-
1.73	+	-	+	+
2.10	-	+	+	-

TABLE IV. QTL contributions to three quantitative traits (growing degree day to anthesis (GDDTA), growing degree day to silking (GDDTS) and plant heightbp) in the four maize lines B73, B97, CML254, and Ki14 estimated from [26].

Our second example is based on QTL data for photoperiod response traits of four different maize strains. The photoperiod response of a trait is defined as the trait difference observed between specimens grown in an environment with long days and specimens grown in a short-day environment. We consider the traits 'days to anthesis' (time from planting to full flower development) and 'days to silking' (silk emergence in maize), both measured in growing degree days (daily average temperature above a threshold temperature of 10 °C cumulated over days of growing). As a third trait we consider plant height. For maize it has been shown that the architecture of quantitative traits such as flowering time and leaf size follows very accurately a model with additive trait effects and only weak epistatic effects [31, 60], as assumed in our model. In [26], the trait contribution of alleles from different QTL is given. As for *Mimulus* above, most of the loci show two alleles with very similar trait contributions and one allele with a significantly higher or lower trait contribution. Yet about one third of the loci have an unclear assignment of alleles (e.g. it is unclear whether the allele of a locus of line 1 is more similar to the allele of line 2 or to allele of line 3) or have more than two significantly different allelic effects. This indicates uncertainties in the allelic effects but may also show limitations of the two-allele model. We exclude from the analysis loci with unclear trait contribution or more than 2 significantly different allelic effects. For the two-allele loci, we estimate the trait contribution g_i of a given locus from the mean of the absolute values of the trait contributions of the different lines for this locus. The resulting values for q_i and g_i are in Table IV.

Two of the lines in [26] (B73, B97) are taken from temperate climates featuring long days in summer and short days in winter, while the other two (CML254, Ki14) are taken from tropical environments with constant length of day over the year. Thus we use as the simplest evolutionary scenario Q_4 with ($Ns_{\rm B73} = -Ns$, $Ns_{\rm B97} = -Ns$, $Ns_{\rm CML254} = +Ns$, $Ns_{\rm Ki14} = +Ns$), with only a single free parameter. We compare this selective scenario Q_4 against the null models P_s and P_0 from (2) with $n = 4$. Again, we apply both the conditioned allele distribution (6) as well as the Holm-Bonferroni correction.

We first consider the 'growing degree day to anthesis' (GDDTA) trait, which measures the time to full flower development. For tropical lines, which are not adapted to long day length, the flowering time is reduced for specimens grown in temperate latitudes compared to tropical environments [26]. For the temperate lines no difference in flowering time is observed between the different environments. For this trait 4 out of 7 loci show a clear two-allele pattern. We first apply the conditioned allele distribution. The null models P_0 and P_s are conditioned on R_{32} for this trait. In this example, the straightforward maximum likelihood estimate of the parameter h fails, since all alleles in the high line are +-alleles and all alleles in the low line are −-alleles, leading to a diverging $h \to \infty$. We use a lower-bound estimate for h by determining the value h for which the probability to see this extreme configuration equals p_l. Here, we choose $p_l = 0.1$ to obtain a conservative estimate for h. For consistency, Ns is determined in the same way. First, we test against the neutral null hypothesis P_0. The log-odds score (3) then gives $S_{Q_4,P_0} = 1.73$ ($p = 0.016$) in favour of the selective scenario. The test for lineage-specific selection yields $S_{Q_4,P_s} = 1.29$ ($p = 0.026$) favouring a lineage-specific scenario. The Holm-Bonferroni correction yields consistent significant results in both cases ($S_{Q_4,P_0} = 4.43$, $p = 0.0023$ and $S_{Q_4,P_s} = 4.09$, $p = 0.0082$). We test the robustness of this result by calculating the number of hidden traits m_h for which these results become insignificant. We obtain $m_h = 17$ and $m_h = 2$, respectively, such that the test against neutrality is more robust than the test against uniform selection strength.

For the 'growing degree day to silking' (GDDTS) trait, with 4 two-allele loci out of 6, the score under conditioning $S_{Q_4,P_0} = 2.05$ ($p = 0.013$) favours the selective scenario over the neutral null model as well. Again we condition on R_{32} and use the lower bound for h described above. Testing for lineage-specific selection under conditioning gives $S_{Q_4,P_s} = 2.04$ ($p = 0.0037$), favouring lineage-specific selection. Using Holm-Bonferroni we obtain a similar result ($S_{Q_4,P_0} = 4.02$, $p = 0.0059$, $m_h = 4$ and $S_{Q_4,P_s} = 4.01$, $p = 0.0082$, $m_h = 3$). Since these two traits show the largest difference in photoperiodic response between tropical and temperate lines, this is a sensible result. The 'plant height' trait, with 4 two-allele loci out of 6, yields no significant score ($S_{Q_4,P_0} = -0.09$, $p = 0.41$) under conditioning and $S_{Q_4,P_0} = 0.13$, $p = 0.044$ in favour of the neutral model with Holm-Bonferroni. Here, h was again determined by maximum likelihood and the conditioning was on R_{34}. So for this trait one cannot reject the hypothesis that the trait difference in the lines is due to neutral evolution. The test for lineage-specific selection also yields no significant result ($S_{Q_4,P_s} = -0.35$, $p = 0.50$ under conditioning and $S_{Q_4,P_s} = -0.13$, $p = 0.21$ with Holm-Bonferroni). The other traits investigated in the study [26] (GDDASI, ear height and total leaf number) have fewer two-allele loci (≤ 3) and none of these traits show a significant support for either of the two hypotheses (data not shown).

Again we also apply Orr's test for comparison. We compare the two lines B73 and CML254, which show the largest phenotypic difference in both the GDDTA and the GDDTS train. For the GDDTA trait 6 out of 6 diverged loci have the +-allele. The p-value $p = 0.13$ is not significant in this case. For the GDDTS trait, 5 out of 5 diverged loci go in the + direction. But also here, Orr's test yields no significant result ($p = 0.2$). A summary of the results can again be found in Table V.

All these results are based on a scoring function that assumes the distribution of + and −-alleles such as (2) is in equilibrium, which is attained at long times after the last common ancestor of the lines. However, the maize lines diverged only around $\lesssim 10000$ years ago [61]. Yet, both in the *Mimulus* and the maize case studies, there are loci that mutated in more than one line from the ancestral state [25, 26] and these mutations affect the trait under consideration, pointing to large mutational targets, which would decrease the time needed to equilibrate the system.

CONCLUSIONS

In this paper, we developed a statistical test to quantify the evidence for different evolutionary scenarios from QTL data for an arbitrary number of lines. The scenarios we consider are neutral evolution, uniform selection strength on the trait across all lines, and lineage-specific selection. Kimura-Ohta theory was used to derive the allele statistics under the different selective scenarios. We find that the use of more than two lines not only increases the statistical power of selection tests, but also their scope: From more than two lines it is possible to infer not only relative selection differences between lines, but also absolute selection levels acting on each line. We applied the test to QTL data on floral characters in different *Mimulus* species and photoperiod response traits in maize, finding signs of lineage-specific selection that were not detectable with a two-line test.

A major bottleneck of the multiple-line test is the need for experimental crosses between three or more different lines. Due to the additional experimental work involved, there are currently relatively few datasets on QTL and their contributions to a trait in more lines than two. However recent studies, employing crosses of 25 maize lines and detecting around 30-40 QTL per trait give a promising outlook into the future [31, 60]. Applying our test to very large numbers of lines poses interesting challenges in connection with the number of alleles per locus and the rapid growth of the number of possible evolutionary scenarios with the number of QTL lines.

A possible application of this test could be inference of gene expression adaptation using expression QTL

Mimulus study	Ns_3	S	p
corolla width			
Q_3 vs. P_0 cond.	1.23	1.05	0.034
Q_3 vs. P_0 H-B	1.23	3.55	0.014
Q_3 vs. P_s cond.	1.23	1.51	0.031
Q_3 vs. P_s H-B	1.23	3.03	0.018
corolla length			
Q_3 vs. P_0 cond.	0.96	-0.53	0.30
Q_3 vs. P_0 H-B	0.96	1.84	0.094
Q_3 vs. P_s cond.	0.96	-0.54	0.36
Q_3 vs. P_s H-B	0.96	1.84	0.048

Maize study	Ns_4	S	p
GDDTA			
Q_4 vs. P_0 cond.	0.94	1.73	0.016
Q_4 vs. P_0 H-B	0.94	4.43	0.0023
Q_4 vs. P_s cond.	0.94	1.29	0.026
Q_4 vs. P_s H-B	0.94	4.09	0.0082
GDDTS			
Q_4 vs. P_0 cond.	0.91	2.05	0.013
Q_4 vs. P_0 H-B	0.91	4.02	0.0059
Q_4 vs. P_s cond.	0.91	2.04	0.0037
Q_4 vs. P_s H-B	0.91	4.01	0.0082
plant height			
Q_4 vs. P_0 cond.	0.75	-0.09	0.41
Q_4 vs. P_0 H-B	0.75	0.13	0.044
Q_4 vs. P_s cond.	0.75	-0.35	0.5
Q_4 vs. P_s H-B	0.75	-0.13	0.21

TABLE V. Summarized results of the selection test on the two datasets [25, 26]. Different evolutionary scenarios are tested against each other using both the conditioned version of the test (cond.) as well as the Holm-Bonferroni correction (H-B). Ns_3 and Ns_4 denote the inferred selection coefficients of the Q scenarios, S and p are the score obtained and the corresponding p-value.

(eQTL) [30]. Since the number of eQTL is typically small for a single gene, the test could be applied on gene modules, e.g. genes belonging to the same pathway or protein complex, allowing to infer selection on individual pathways. Another future perspective for this method may arise if genome-wide association studies (GWAS) with fully sequenced organisms enable the inference of causal mutations behind the QTL [7, 62], allowing to apply multiple-line tests without the need to perform crosses between different lines.

APPENDIX I: SHORT-TIME DYNAMICS

For completeness, we also discuss the limit of short evolutionary times, where each locus has undergone at most one mutation across the different lines. Since we are interested in loci differing in state across the different lines, this means each locus of interest has undergone exactly one mutation in some line since the last common ancestor. Any statistical test based on the allele statistics at different loci requires a number of loci differing by state. Short evolutionary times are characterized by $\mu t \ll 1$, nevertheless the total number of diverged loci, characterized by $\mu t n$ (where n is the total number of loci) must still be at least of order one.

At short evolutionary times, the probabilities for the alleles $q_1, q_2, \ldots$ also depend on the ancestral state of the locus, denoted $c = \pm 1$. In addition, the phylogenetic tree of the lines has to be considered, since not all final configurations for a diverged locus can be reached by a single mutation from a given ancestral state (see Figure 9). This makes it difficult to construct a general n-line selection model in this limit. Here, we discuss the case of two and three lines.

We start with the case of two lines. The short time transition rates under selection for a given locus in a given line are $\frac{2Ns_a g\mu}{1-e^{-2Ns_a g}}$ and $\frac{-2Ns_a g\mu}{1-e^{2Ns_a g}}$ for the transition from $c = -$ to $q_a = +$ and $c = +$ to $q_a = -$, respectively [38]. Here, μ

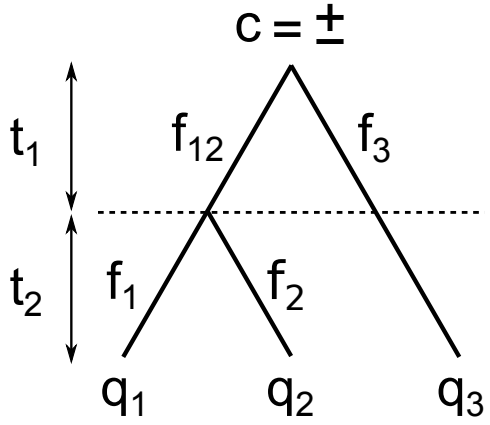

FIG. 9. **Phylogenetic tree for three lines.** In the short-time limit, the allele configurations (q_1, q_2, q_3) which can be reached by a single mutation from an ancestral state c depend on the phylogenetic tree. The branch lengths t_1 and t_2 determine the relative mutation probabilities in the different branches.

is the mutation rate per locus. This leads to probabilities

$$P(a|g,c) = \frac{s_c(g, Ns_a)}{s_c(g, Ns_1) + s_c(g, Ns_2)} \tag{9}$$

for the two diverged states $(+-)$ and $(-+)$, where $s_c(g, Ns) = \frac{-2Nscg}{1-e^{2Nscg}}$ and Ns_a is the selection strength of the diverged line. Given two lines, both final configurations $(q_1, q_2) = (+-)$ and $(-+)$ can be reached from either ancestor $c = \pm$. If the ancestral states are unknown we can average over both possible ancestors, writing

$$P(q_1, q_2|g) = \frac{P(q_2)s_{q_2}(g, Ns_1) + P(q_1)s_{q_1}(g, Ns_2)}{\sum\limits_{c=\pm 1}\sum\limits_{i=1}^{2} P(c)s_c(g, Ns_i)} \tag{10}$$

where $P(c)$ denotes the prior probability to have ancestral states c at the locus in question. The ancestor could have been under selection as well, causing an biased distribution of ancestral alleles $c = \pm$. We write the distribution of ancestral alleles as $P(c) = e^{-cNs_{\rm anc}g}$ with an additional selection parameter $Ns_{\rm anc}$ for the ancestral line. Here, we have assumed that the distribution of ancestors has reached equilibrium.

Considering three lines, four of the six possible diverged configurations can be assigned a unique ancestor: Denoting line 3 as the outgroup line (see Figure 9), configurations $(q_1, q_2, q_3) = (+--)$ and $(-+-)$ diverged from the ancestral allele $c = -$, configurations $(-++)$ and $(+-+)$ from ancestor $c = +$. Configurations $(++-)$ and $(--+)$ can either be reached by a mutation in the ancestor of lines 1 and 2 or a mutation in line 3. One can write the relative probabilities of the 6 allelic configurations excluding $q_1 = q_2 = q_3$ as

$$\begin{aligned}
--+ \quad & (t_1+t_2)P(-)s_-(g, Ns_3) + t_1P(+)s_+(g, Ns_{12}) \\
-+- \quad & t_2P(-)s_-(g, Ns_2) \\
+-- \quad & t_2P(-)s_-(g, Ns_1) \\
++- \quad & t_1P(-)s_-(g, Ns_{12}) + (t_1+t_2)P(+)s_+(g, Ns_3) \\
+-+ \quad & t_2P(+)s_+(g, Ns_2) \\
-++ \quad & t_2P(+)s_+(g, Ns_1)\,,
\end{aligned} \tag{11}$$

where we have a different selection strength s_i in the Kimura transition rates $s_c(g, Ns)$ for each line in the most general case. The times t_1 and t_2 account for the different branch lengths of the phylogenetic tree (see Figure 9). The mutation rate μ drops out in (11) due to normalization. This result is written down in a more compact form in eq. (4) in main text.

Again, allele statistics resulting from the different hypotheses P_0, P_s and Q_1 etc. can be tested against each other using log-odds score as in (3). We perform numerical simulations analogous to the simulations performed for the

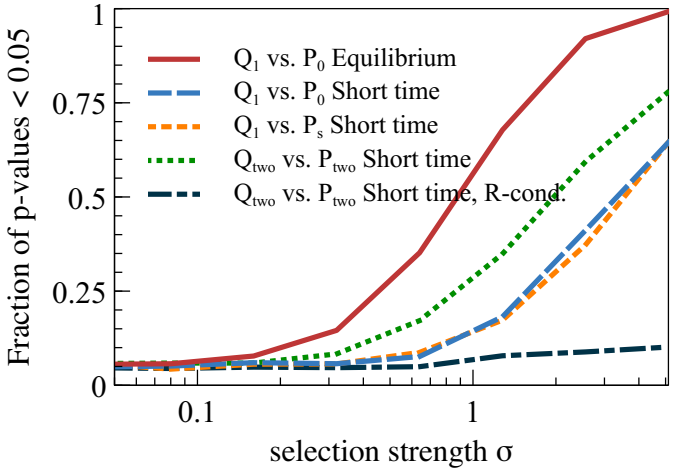

FIG. 10. **Statistical significance of the tests for short evolutionary times.** The selection test for short evolutionary times applied to artificial data created in the short-time limit shows a little less statistical power compared to the equilibrium test applied to data generated at long evolutionary times, but still allows to identify selection in a reasonable parameter range. However, for two lines under conditioning on $R_{\max}$ the short-time test barely has any statistical power, analogously to the equilibrium case, where it has none. Simulation parameters: The phylogenetic branch lengths t_1 and t_2 are taken equal to each other. Ancestral states c of the loci assigned with probability $P(c) = 1/2$. The other simulation parameters are as in Figure 4.

equilibrium test, setting $t_1 = t_2$. In all cases we do not assume knowledge of the ancestral states but estimate the selection parameter Ns_{anc} of the ancestral line via maximum likelihood together with the other selection parameters to determine $P(c)$. Under the Q_1 selective scenario the simulations show that the statistical power of the short-time test on three lines on short-time data is lower than the three-line equilibrium test applied to data for long evolutionary times (see Figure 10), but still allows to detect selection in the short time case. For two lines the test under conditioning on $R_{\max}$ again gives hardly any significant results for short times (see Figure 10), while the $R_{\max}$-conditioning for three lines as well as the Holm-Bonferroni correction for two and three lines result in functioning selection tests.

APPENDIX II: PEDAGOGICAL EXAMPLE FOR THE MAXIMUM ENTROPY PRINCIPLE

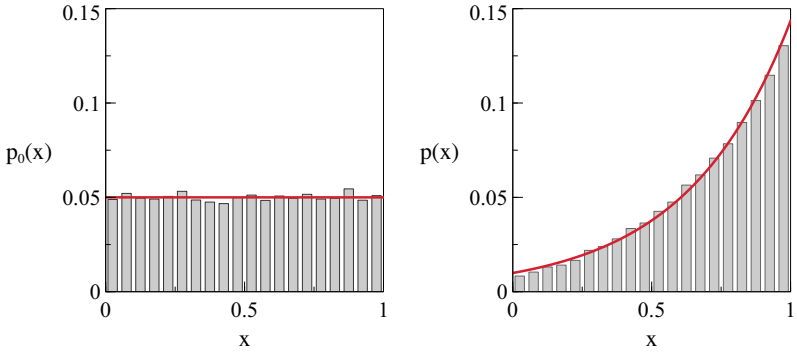

FIG. 11. **A biased distribution can be inferred with the maximum entropy method.** Ascertainment leads to a biased distribution, which is derived using the maximum entropy method. Left: Histogram for 1000 sets of ten random numbers each drawn from a uniform distribution (red line) in the interval [0,1]. Rightbp: Only sets of numbers are retained which have a sum S close to $m = 8$ ($7.95 < S < 8.05$). In these sets higher numbers appear more often than in the uniform distribution. The biased distribution takes on an exponential form given by the maximum-entropy distribution (15) (red line).

Here, we give a simple concrete example to illustrate the link between ascertainment bias and the maximum entropy principle. Consider a uniform distribution $p(x)$ on the interval $[0, 1]$, from which ten numbers are drawn independently

(see fig. 11 left). If one repeatedly draws such sets of ten numbers, the sum over each set will fluctuate from set to set with a mean value of 5. In the next step, we only retain those sets whose sum is close to some value of $m \neq 5$. The numbers in these sets follow a non-uniform distribution and for $m > 5$ one finds that larger values x appear with a higher probability compared to the uniform distribution (see fig. 11 right). Although each of these numbers was originally drawn from the uniform distribution, retention of sets with a particular mean value introduces a bias in the observed distribution of x. This is the ascertainment bias induced by conditioning the sum of each set. The principle of maximum entropy allows to determine the exact form of this biased distribution $p(x)$. We maximize the relative information entropy between the distribution $p(x)$ and the original (uniform) distribution $p_0(x) = 1$ for $x \in [0, 1]$

$$H(p) = -\int_0^1 \mathrm{d}x\, p(x) \log \frac{p(x)}{p_0(x)}\,, \tag{12}$$

subject to the constraints

$$\int_0^1 \mathrm{d}x\, p(x) = 1\,, \qquad \int_0^1 \mathrm{d}x\, xp(x) = \frac{m}{N}\,, \tag{13}$$

where $N = 10$ is the size of each set. Here, the first constraint ensures the normalization of $p(x)$ and the second constraint fixes the mean value of x to m/N. Introducing Lagrange multipliers to maximize (12) subject to the constraints (13) leads to [50]

$$\begin{aligned} &-\int_0^1 \mathrm{d}x\, p(x) \log p(x) + \lambda_1 \left(\int_0^1 \mathrm{d}x\, p(x) - 1 \right) \\ &+ \lambda_2 \left(\int_0^1 \mathrm{d}x\, xp(x) - m/N \right) \end{aligned} \tag{14}$$

to be maximized with respect to $p(x)$. Differentiating (14) with respect to p and setting the derivative to zero gives

$$p(x) = e^{\lambda_2 x + \lambda_1 - 1}. \tag{15}$$

Ascertainment bias thus makes x exponentially rather than uniformly distributed, with coefficients λ_1 and λ_2 determined by the constraints (13). For $m = 8$ and $N = 10$ one obtains $\lambda_1 \approx -1.62$ and $\lambda_2 \approx 2.67$; the result for $p(x)$ shown in fig. 11 agrees perfectly with the histogram of numbers in sets with a constrained sum.

Suppose one did not know whether the original distribution $p(x)$ from which the data were drawn was uniform or not and one had access only to data subject to the known constraint. If the distribution of the empirical data deviates from or agrees with the maximum entropy distribution $p(x)$, then this deviation or agreement could be used to quantify the likelihood that the original data came from the uniform distribution (vs. an alternative hypothesis). We follow the analogous approach with the score (7) to tell whether a particular allele statistics more likely comes from neutral evolution in combination with ascertainment bias (vs. an alternative scenario involving selection).

ACKNOWLEDGMENTS

We gratefully acknowledge discussions with Andreas Beyer and Mathieu Clément-Ziza. This work was supported by the DFG under SFB 680.

[1] C. L. Dilda and T. F. C. Mackay, Genetics **162**, 1655 (2002).
[2] J. G. Mezey, D. Houle, and S. V. Nuzhdin, Genetics **169**, 2101 (2005).
[3] S. V. Nuzhdin, A. A. Khazaeli, and J. W. Curtsinger, Genetics **170**, 719 (2005).
[4] T. F. Mackay and R. F. Lyman, Philos. T. Roy. Soc. B **360**, 1513 (2005).
[5] R. B. Brem and L. Kruglyak, Proc. Natl. Acad. Sci. USA **102**, 1572 (2005).

[6] J. Flint and T. F. Mackay, Genome Res. **19**, 723 (2009).
[7] T. F. C. Mackay, E. A. Stone, and J. F. Ayroles, Nat. Rev. Genet. **10**, 565 (2009).
[8] E. G. Pasyukova, C. Vieira, and T. F. C. Mackay, Genetics **156**, 1129 (2000).
[9] J. J. Fanara, K. O. Robinson, S. M. Rollmann, R. R. H. Anholt, and T. F. C. Mackay, Genetics **162**, 1321 (2002).
[10] M. De Luca, N. V. Roshina, G. L. Geiger-Thornsberry, R. F. Lyman, E. G. Pasyukova, and T. F. C. Mackay, Nat. Genet. **34**, 429 (2003).
[11] A. J. Moehring and T. F. C. Mackay, Genetics **167**, 1249 (2004).
[12] S. T. Harbison, A. H. Yamamoto, J. J. Fanara, K. K. Norga, and T. F. C. Mackay, Genetics **166**, 1807 (2004).
[13] K. W. Jordan, T. J. Morgan, and T. F. C. Mackay, Genetics **174**, 271 (2006).
[14] A. Rebai and B. Goffinet, Theor. Appl. Genet. **86**, 1014 (1993).
[15] J. Steinhoff, W. Liu, H. P. Maurer, T. Wrschum, H. L. C. Friedrich, N. Ranc, and J. C. Reif, Crop Sci. **51**, 2505 (2011).
[16] G. Blanc, A. Charcosset, B. Mangin, A. Gallais, and L. Moreau, Theor. Appl. Genet. **113**, 206 (2006).
[17] J.-L. Jannink and R. Jansen, Genetics **157**, 445 (2001).
[18] C. Rückert and J. Bennewitz, Genet. Sel. Evol. **42**, 40 (2010).
[19] S. Crepieux, C. Lebreton, B. Servin, and G. Charmet, Genetics **168**, 1737 (2004).
[20] K. J. F. Verhoeven, J.-L. Jannink, and L. M. McIntyre, Heredity **96**, 139 (2006).
[21] C. Xie, D. D. G. Gessler, and S. Xu, Genetics **149**, 1139 (1998).
[22] A. Rebai and B. Goffinet, Genet. Res. **75**, 243 (2000).
[23] S. Xu, Genetics **148**, 517 (1998).
[24] N. Yi and S. Xu, Genetica **114**, 217 (2002).
[25] C. Chen, *Lineage specific inference about QTL evolution among three Mimulus species of contrasting relationship and inbreeding*, Ph.D. thesis, University of British Columbia (2009).
[26] N. D. Coles, M. D. McMullen, P. J. Balint-Kurti, R. C. Pratt, and J. B. Holland, Genetics **184**, 799 (2010).
[27] H. A. Orr, Genetics **149**, 2099 (1998).
[28] D. P. Rice and J. P. Townsend, Genetics **190**, 1533 (2012a).
[29] H. B. Fraser, A. M. Moses, and E. E. Schadt, Proc. Natl. Acad. Sci. USA **107**, 2977 (2010).
[30] H. B. Fraser, Bioessays **33**, 469 (2011).
[31] E. S. Buckler, J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown, C. Browne, E. Ersoz, S. Flint-Garcia, A. Garcia, J. C. Glaubitz, M. M. Goodman, C. Harjes, K. Guill, D. E. Kroon, S. Larsson, N. K. Lepak, H. Li, S. E. Mitchell, G. Pressoir, J. A. Peiffer, M. O. Rosas, T. R. Rocheford, M. C. Romay, S. Romero, S. Salvo, H. S. Villeda, H. Sofia da Silva, Q. Sun, F. Tian, N. Upadyayula, D. Ware, H. Yates, J. Yu, Z. Zhang, S. Kresovich, and M. D. McMullen, Science **325**, 714 (2009).
[32] N. H. Barton and H. P. de Vladar, Genetics **181**, 997 (March 2009).
[33] H. P. de Vladar and N. H. Barton, Trends Ecol. Evol. **26**, 424 (2011).
[34] A. Nourmohammad, T. Held, and M. Lässig, Curr. Opin. Genet. Dev. **23**, 684 (2013a).
[35] A. Nourmohammad, S. Schiffels, and M. Lässig, J. Stat. Mech. - Theory E. **2013** (2013b).
[36] R. Lande, Heredity **50**, 47 (February 1983).
[37] L.-M. Chevin and F. Hospital, Genetics **180**, 1645 (November 2008).
[38] M. Kimura and T. Ohta, Genetics **61**, 763 (1969).
[39] Y. Iwasa, J. Theor. Biol. **135**, 265 (1988).
[40] J. Berg, S. Willmann, and M. Lässig, BMC Evol. Biol. **4**, 42 (2004).
[41] G. Sella and A. E. Hirsh, Proc. Natl. Acad. Sci. USA **102**, 9541 (2005).
[42] G. Schwarz, Ann. Stat. **6**, 461 (1978).
[43] Z. B. Zeng, Genetics **131**, 987 (1992).
[44] R. D. Bickel, A. Kopp, and S. V. Nuzhdin, PLoS Genet. **7**, e1001275 (2011).
[45] M. G. Kidwell and D. Lisch, Proc. Natl. Acad. Sci. USA **94**, 7704 (1997).
[46] C. Feschotte, N. Jiang, and S. R. Wessler, Nature **3**, 329 (1997).
[47] K. Naito, F. Zhang, T. Tsukiyama, H. Saito, C. N. Hancock, A. O. Richardson, Y. Okumoto, T. Tanisaka, and S. R. Wessler, Nature **461**, 1130 (2009).
[48] A. Studer, Q. Zhao, J. Ross-Ibarra, and J. Doebley, Nat. Genet. **43**, 1160 (2011).
[49] H. Fu and H. K. Dooner, Proc. Natl. Acad. Sci. USA **99**, 9573 (2002).
[50] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[51] R. Narayan and R. Nityananda, Annu. Rev. Astron. Astr. **24**, 127 (1986).
[52] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, Comput. Linguist. **22**, 39 (1996).
[53] A. Prügel-Bennett and J. L. Shapiro, Phys. Rev. Lett. **72**, 1305 (1994).
[54] A. Prügel-Bennett and J. L. Shapiro, Physica D **104**, 75 (1997).
[55] M. Ruttray, Complex Syst. **9**, 213 (1995).
[56] V. Mustonen and M. Lässig, Proc. Natl. Acad. Sci. USA **102**, 15936 (2005).
[57] M. Lässig, BMC Bioinformatics **8**, S7 (2007).
[58] V. Mustonen, J. Kinney, C. G. Callan, and M. Lässig, Proc. Natl. Acad. Sci. USA **105**, 12376 (2008).
[59] A key difference of log-odds score to Orr's test is that Orr not only uses the empirically observed phenotypic effects g_i available from crossing experiments, but also phenotypic effects drawn from plausible distribution $P(g)$. Previous work by Rice and Townsend found that the outcome of Orr's test strongly depends on the assumptions made on this distribution and that the test can produce nonsensical results [63] in particular cases. For instance, Rice and Townsend showed that the

variance of the phenotypic effects strongly affects the test outcome and that for a vanishing trait effect variance the test never rejects neutrality, even when all loci have +-alleles. Additionally, they observed that when selection strength acting on a locus depends on the trait contributions g_i of that locus the test rejected the null hypothesis of neutrality more often for a neutral scenario ($s = 0$) than for a scenario with a positive selection strength acting on one of the lines ($s > 0$). Here, we showed that based only on the trait contributions g_i found experimentally (making no assumptions about a possible distribution of unobserved trait effects), a two-line selection test under Orr's conditioning always yields a nil result in equilibrium. At short evolutionary times we find that the log-odds scores (7) are numerically small, see appendix I.

[60] F. Tian, P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun, S. Flint-Garcia, T. R. Rocheford, M. D. McMullen, J. B. Holland, and E. S. Buckler, Nat. Genet. **43**, 159 (2011).

[61] J. Doebley, Annu. Rev. Genet. **38**, 37 (2004).

[62] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher, Nature **461**, 747 (2009).

[63] D. P. Rice and J. P. Townsend, G3 **2**, 905 (2012b).