The Annals of Applied Statistics
2020, Vol. 14, No. 1, 202–220
https://doi.org/10.1214/19-AOAS1301

MODIFYING THE CHI-SQUARE AND THE CMH TEST FOR POPULATION GENETIC INFERENCE: ADAPTING TO OVERDISPERSION

BY KERSTIN SPITZER[1,*], MARTA PELIZZOLA[1,†] AND ANDREAS FUTSCHIK[2]

[1]*Vienna Graduate School of Population Genetics, Vetmeduni Vienna,* [*]*kerstin.e.spitzer@gmail.com;* [†]*Marta.Pelizzola@vetmeduni.ac.at*

[2]*Department of Applied Statistics, Johannes Kepler University Linz, andreas.futschik@jku.at*

Evolve and resequence studies provide a popular approach to simulate evolution in the lab and explore its genetic basis. In this context, Pearson's chi-square test, Fisher's exact test as well as the Cochran–Mantel–Haenszel test are commonly used to infer genomic positions affected by selection from temporal changes in allele frequency. However, the null model associated with these tests does not match the null hypothesis of actual interest. Indeed, due to genetic drift and possibly other additional noise components such as pool sequencing, the null variance in the data can be substantially larger than accounted for by these common test statistics. This leads to p-values that are systematically too small and, therefore, a huge number of false positive results. Even, if the ranking rather than the actual p-values is of interest, a naive application of the mentioned tests will give misleading results, as the amount of overdispersion varies from locus to locus. We therefore propose adjusted statistics that take the overdispersion into account while keeping the formulas simple. This is particularly useful in genome-wide applications, where millions of SNPs can be handled with little computational effort. We then apply the adapted test statistics to real data from *Drosophila* and investigate how information from intermediate generations can be included when available. We also discuss further applications such as genome-wide association studies based on pool sequencing data and tests for local adaptation.

1. Introduction. An important question in the field of population genetics is how populations adapt to changes in their environment. Experimental evolution allows to study the adaptation under controlled conditions. If these experiments are combined with high-throughput sequencing, they are called evolve and resequence (E&R) experiments (Turner et al. (2011)). Such experiments are carried out both on microbes and on higher organisms. Due to large population sizes and short generation times, microbes permit to study evolutionary processes based on newly arriving mutations. With higher organisms and sexual reproduction on the other hand, evolution based on standing genetic variation is usually explored. A major goal is to identify genomic positions (a.k.a. loci) that are responsible for the adaptation. For this purpose, the organisms are kept over $t + 1$ generations $G_0, G_1, \ldots, G_t$ under conditions that require adaptation. Allele frequencies are obtained by sequencing the genomes at G_0 and G_t and possibly also at some intermediate time points. Depending on the (time and financial) budget, the organisms are sequenced individually, or as a pool, in order to obtain allele frequencies for typically millions of single-nucleotide polymorphisms (SNPs). Individual sequencing can also be implemented using barcoding, with barcode tags that identify the organism being added before sequencing.

Allele frequency changes over time are then tested for signals of selection. For this purpose, usually only biallelic SNPs are considered. Indeed, multiallelic sites are rare in population data and likely caused by sequencing errors (Burke et al. (2010)). Consequently, frequencies of the two alleles are used in the base and the evolved population for each tested SNP.

Received February 2019; revised August 2019.
Key words and phrases. Chi-square test, CMH test, overdispersion, experimental evolution, evolve and resequence, genetic drift, pool sequencing.

Pearson's chi-square test (for simplicity subsequently called chi-square test) is a very popular method for this purpose (e.g., Griffin et al. (2017)). Serving the same purpose, Fisher's exact test is sometimes used as an alternative (Burke et al. (2010)). Being a generalization of the chi-square test for stratified data, the Cochran–Mantel–Haenszel (CMH) test is also often applied when allele frequency data from several replicate populations is available. See, for example, Barghi et al. (2017), Nouhaud et al. (2016), Orozco-terWengel et al. (2012), Tobler et al. (2014) for applications.

Kofler and Schlötterer (2014) compare several methods for detecting selection by contrasting allele frequencies at two different time points. Apart from the CMH test, they consider the pairwise summary statistic "diffStat" (Turner et al. (2011)), an association statistic by Turner and Miller (2012) and F_{ST} (Remolina et al. (2012)). A comparison of receiver operator curves (ROC) for these tests shows that the CMH test performs best, that is, has more power than the other tests considered to identify selected SNPs.

Further methods are available for detecting selection in E&R experiments when organisms are sequenced also at intermediate generations. The method of Bollback, York and Nielsen (2008), for instance, is based on a hidden Markov model (HMM). Generalizations of this approach are provided by Malaspinas et al. (2012) and Steinrücken, Bhaskar and Song (2014). Also, Mathieson and McVean (2013) adapt HMMs to structured populations. The method CLEAR by Iranmehr et al. (2017) uses Markov chains in a discrete state model and computes the exact likelihood for small populations. Ignoring spatial dependence, linked loci are modeled using composite likelihood statistics. A frequency increment test (FIT) based on an approximation of the allele frequency dynamics by a Gaussian process is proposed by Feder, Kryazhimskiy and Plotkin (2014). Considering all loci separately, Topa et al. (2015) also model the allele frequency trajectories by Gaussian processes, whereas Terhorst, Schlötterer and Song (2015) approximate the joint likelihood for multiple loci. The approach by Taus, Futschik and Schlötterer (2017) is based on linear least square (LLS) regression to fit the allele frequency data to a selection model. Schraiber, Evans and Slatkin (2016) estimate parameters in a Bayesian framework with Markov chain Monte Carlo sampling. Another Bayesian approach has been proposed by Levy et al. (2015) for estimating parameters in barcoded lineages. Finally, the Wright–Fisher ABC method proposed by Foll, Shim and Jensen (2015) applies approximate Bayesian computation (ABC).

Besides detecting loci under selection, several of the discussed approaches additionally estimate selection coefficients, often jointly, with other parameters like the effective population size (e.g., Bollback, York and Nielsen (2008)), age of alleles (e.g., Malaspinas et al. (2012), Schraiber, Evans and Slatkin (2016)) or allelic dominance (e.g., Taus, Futschik and Schlötterer (2017)). However, such methods are computationally much more demanding than simple methods like the chi-square and the CMH tests. Hence, the latter are still widely used when testing for selection and, for example, implemented in the software tool PoPoolation2 (Kofler, Pandey and Schlötterer (2011)). A recent comparison of several methods to detect different types of selection by Vlachos et al. (2019) also suggests that, besides their simplicity, the adapted CMH and chi-square tests proposed here are among the best performing methods.

When comparing the allele frequencies between pairs of samples, the null models for the classical chi-square test, Fisher's exact test and also the CMH test assume that the probability of sampling a given allele is the same within each given pair. However, the sampling variation is not the only component of variance relevant in E&R experiments. Allele frequency changes between generations happen because of genetic drift, that is, due to chance. This increases the variance in the data noticeably, unless population sizes are large enough to safely ignore drift. Such a situation usually occurs only when working with microorganisms (e.g., Illingworth et al. (2012)). Another potential source of random variation is pool sequencing where the obtained reads can be regarded as a sample from the DNA pool.

If the chi-square or the CMH tests are applied to data, which contain more variance than assumed by the tests (overdispersion), the resulting values of the test statistic will be too large and, hence, the p-values too small. As a consequence, selection is often inferred for loci where it is not present. In the simulations with drift and pool sequencing that we carried out, the null hypothesis of neutrality is rejected in up to 80% of the cases, despite being true. Hence, the additional variance introduced by drift and pool sequencing is by no means negligible. A common procedure to account for false positives due to drift and pool sequencing is to calculate a modified rejection cutoff via computer simulations (Orozco-terWengel et al. (2012)). Griffin et al. (2017) chose another approach by applying three different statistical tests and considering only SNPs which are significant with respect to all three methods as candidate loci.

This issue is also well known in the unrelated context of complex surveys, where different strategies have been proposed to obtain appropriate tests of homogeneity. (See, e.g., Chapter 10 in Lohr (2010).)

Our aim is to adapt the chi-square test and the CMH test in a way that additional sources of variance are directly included into the test statistics, making computer simulations no longer necessary. Thus, we propose a method that is faster than the commonly used ones. When sequence data for intermediate generations is available, the additional information can be included into our test statistics without a considerable increase in computation time. Also in terms of power, our method performs better than other approaches. In particular, our method has considerably more power to detect selected SNPs than the classical CMH test with the modified rejection cutoff (Orozco-terWengel et al. (2012)). Compared to the approach for detecting selection in Taus, Futschik and Schlötterer (2017), which the authors found to be faster than the CLEAR method by Iranmehr et al. (2017) and the Wright–Fisher ABC by Foll, Shim and Jensen (2015), our method has also slightly more power and is 10^5 times faster. Hence, our method performs very well in terms of speed and power.

In this article, we first present variants of the chi-square and the CMH tests for general underlying variances. The statistics derived in this step can be useful in all situations, where overdispersion is present.

Further, we provide specific formulas for the test statistics under scenarios with drift and pool sequencing, which are common in E&R experiments, and seen also in other situations. In genome-wide association studies (GWAS), for instance, the CMH test is often used for the inference of an association between a trait such as a disease and an allele variant. When the data arises from pool sequencing (e.g., Bastide et al. (2013), Endler et al. (2016)), our adapted test could be a good alternative. This and further applications of our proposed tests are discussed in Section 5.

The remainder of this article is structured as follows: The test statistics for general underlying variances are presented in Section 2, and the scenarios with drift and pool sequencing are considered in Section 2.1. In Section 3 the adapted tests are applied to simulated data, and their performance is examined. We use the tests on real data and present the results in Section 4. A discussion in Section 5 concludes this article.

2. Adapted tests. In this section, we generalize the chi-square and the CMH tests to work under overdispersion and derive explicit formulas for scenarios with drift and pool sequencing. Given the application in mind, our focus is on the null model of homogeneity, although our derivations also apply to the tests of independence.

We summarize our data in a 2×2 contingency table. Using the notation from Table 1, the chi-square test statistic in its standard form is defined as

$$T_{\chi^2} := \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(x_{ij} - \frac{x_{i+}x_{+j}}{n})^2}{\frac{x_{i+}x_{+j}}{n}} = \frac{n(x_{11}x_{22} - x_{12}x_{21})^2}{x_{1+}\, x_{2+}x_{+1}x_{+2}}. \tag{1}$$

TABLE 1

Standard contingency table used with chi-square test. Subsequent interpretation in our population genetic application: Entries are allele frequencies for a biallelic SNP taken either from two populations or from one population at two time points. n is the total number of sequencing reads (coverage) for the considered SNP, x_{ij} are the reads for allele j in population i, x_{i+} is the total number of reads in population i and x_{+j} is the total number of allele j in both populations, $i, j \in \{1, 2\}$. The frequencies are obtained either by individual sequencing of a sample or by pool sequencing applied to the entire population

	allele 1	allele 2	
Population 1	x_{11}	x_{12}	x_{1+}
Population 2	x_{21}	x_{22}	x_{2+}
	x_{+1}	x_{+2}	n

As shown in Section S1 (equation (S10)) of the Supplementary Material (Spitzer, Pelizzola and Futschik (2020)), we may rewrite T_{χ^2} as

$$T^a_{\chi^2}(\hat{s}_1^2, \hat{s}_2^2) := \frac{(x_{11}x_{22} - x_{12}x_{21})^2}{x_{2+}^2\hat{s}_1^2 + x_{1+}^2\hat{s}_2^2} = \frac{(x_{11} - \frac{x_{1+}x_{+1}}{n})^2}{(\frac{x_{2+}}{n})^2\hat{s}_1^2 + (\frac{x_{1+}}{n})^2\hat{s}_2^2}, \tag{2}$$

with $\hat{s}_1^2 := x_{1+}\frac{x_{+1}}{n}\frac{x_{+2}}{n}$, and $\hat{s}_2^2 := x_{2+}\frac{x_{+1}}{n}\frac{x_{+2}}{n}$.

In order to adapt the test to models that involve different variances, we may now replace $\hat{s}_1^2$ and $\hat{s}_2^2$ by consistent estimators of $\mathrm{Var}(x_{11})$ and $\mathrm{Var}(x_{21})$. As for the classical chi-square test (see Section S1 of the Supplementary Materials (Spitzer, Pelizzola and Futschik (2020)), in particular equation (S8)), $T^a_{\chi^2}(\hat{s}_1^2, \hat{s}_2^2)$ converges in distribution to a χ^2-distribution with one degree of freedom under the null hypothesis of homogeneity.

The CMH test is based on a $2 \times 2 \times k^*$ contingency table, where the k^* partial 2×2 tables are assumed to be independent. We use the same notation as for the chi-square test and indicate a variable belonging to the kth partial table with the additional index k such as in x_{+1k} below. The null hypothesis is that both true proportions within each partial table are the same, that is, the odds ratio in each partial table equals 1 (McDonald (2014)). The classical CMH test statistic is

$$T_{\mathrm{CMH}} := \frac{(\sum_{k=1}^{k^*}(x_{11k} - \frac{x_{1+k}x_{+1k}}{n_k}))^2}{\sum_{k=1}^{k^*}\frac{x_{1+k}x_{+1k}x_{2+k}x_{+2k}}{n_k^2(n_k-1)}}. \tag{3}$$

See Chapter 6.3 in Agresti (2002).[1]

Analogous to the chi-square test, one can adapt the CMH test to general underlying variances. As a first step, we write the test statistic of the CMH test as

$$T^a_{\mathrm{CMH}}(\hat{s}_{1k}^2, \hat{s}_{2k}^2; k = 1, \ldots, k^*) := \frac{(\sum_{k=1}^{k^*}(x_{11k} - \frac{x_{1+k}x_{+1k}}{n_k}))^2}{\sum_{k=1}^{k^*}((\frac{x_{2+k}}{n_k})^2\hat{s}_{1k}^2 + (\frac{x_{1+k}}{n_k})^2\hat{s}_{2k}^2)} \tag{4}$$

and insert $x_{1+k}\frac{x_{+1k}}{n_k}\frac{x_{+2k}}{n_k-1}$ for $\hat{s}_{1k}^2$ and $x_{2+k}\frac{x_{+1k}}{n_k}\frac{x_{+2k}}{n_k-1}$ for $\hat{s}_{2k}^2$, $k = 1, \ldots, k^*$. (See Section S2 of the Supplentary Material (Spitzer, Pelizzola and Futschik (2020)).) As with the chi-square test, the formula assumes one sampling step only, which is not appropriate for more complex models. Again, however, $\hat{s}_{1k}^2$ and $\hat{s}_{2k}^2$ can be replaced by consistent estimators of $\mathrm{Var}(x_{11k})$ and $\mathrm{Var}(x_{21k})$, $k = 1, \ldots, k^*$. In the next section we present suitable variance estimators for situations with drift and pool sequencing.

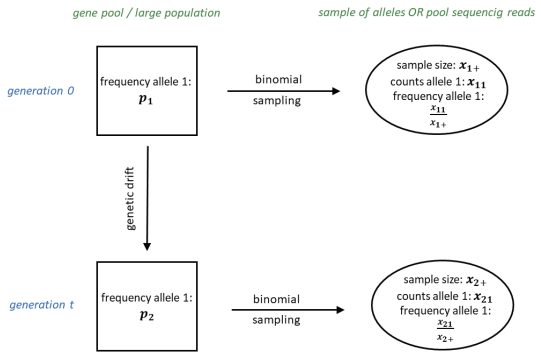
[1] This is the test statistic as proposed by Mantel and Haenszel which is commonly considered for the CMH test. The statistic proposed by Cochran differs by the factor $\frac{1}{n_k}$ instead of $\frac{1}{n_k-1}$ in each term of the denominator. Asymptotically, this difference is negligible.

2.1. *Adaptation of the tests to drift and pool sequencing.* We first focus on the chi-square test and assume that allele frequency data is available for a single population at two time points. Under the null hypothesis, the population allele frequency at the later time point p_2 for a given biallelic SNP arises from p_1 according to the Wright–Fisher model of genetic drift (see, e.g., Chapter 3 in Ewens (2004)). Therefore, p_2 is modeled as a random variable. Usually, one cannot observe p_1 and p_2 directly. Indeed, both quantities are frequently estimated from population samples. If experimenters use pool sequencing as a further sampling step, we model this by binomial sampling as, for example, in Waples (1989) or Jónás et al. (2016). The random sample size is known as coverage, and the success probability is taken as the frequency of allele 1 in the underlying DNA material. If only a sample of the population is sequenced, we assume again binomial sampling for simplicity: An extension to hypergeometric sampling is straightforward.

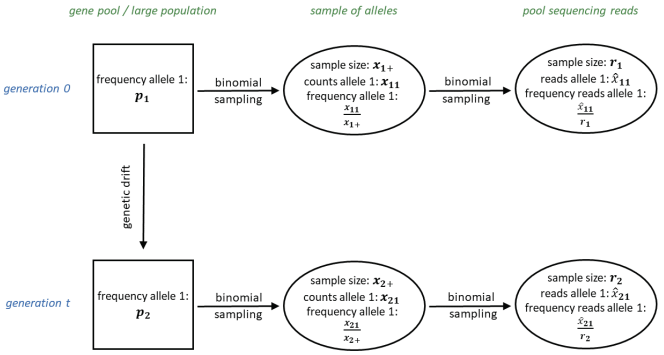
Typically, genomic selection is taken as alternative hypothesis that leads to differences between our estimates for p_1 and p_2 that cannot be explained by sampling and drift.

Figure 1(a) summarizes the scenario with drift and one underlying sampling step, which is either taking a sample from the population for (individual) sequencing or applying pool sequencing (to the whole population). The scenario with drift and two sampling steps (sampling from the population and pool sequencing) is outlined in Figure 1(b).

Table 1 summarizes the notation for a contingency table based on one sampling step (sampling from the population or pool sequencing), while Table 2 is for two sampling steps (sampling plus pool sequencing). Note that in the scenario with two sampling steps, only the



(a) one sampling step



(b) two sampling steps

FIG. 1. *Sampling schemes for the scenarios with genetic drift and one sampling step, shown in* (*a*), *or genetic drift and two sampling steps, shown in* (*b*).

TABLE 2

Allele frequencies for a biallelic SNP taken either from two populations or from one population at two time points, assuming two underlying sampling steps. m is the total number of sequencing reads (coverage) for the considered SNP, $\hat{x}_{ij}$ are the reads for allele j in population i, r_i is the total number of reads in population i and $\hat{x}_{+j}$ is the total number of allele j in both populations, $i, j \in \{1, 2\}$

	allele 1	allele 2	
Population 1	$\hat{x}_{11}$	$\hat{x}_{12}$	r_1
Population 2	$\hat{x}_{21}$	$\hat{x}_{22}$	r_2
	$\hat{x}_{+1}$	$\hat{x}_{+2}$	m

sample sizes x_{1+} and x_{2+}, but not the allele frequencies, are known for the first step. In the situations where one population descends from another, population 1 is the base population, and population 2 is the evolved population.

Conditional on p_1, the variance in allele frequency, due to drift after t generations, can be calculated as

$$p_1(1-p_1)\left(1-\left(1-\frac{1}{2N_e}\right)^t\right), \tag{5}$$

where p_1 is the frequency of allele 1 in the base population and $2N_e$ is the effective population size considering gametes (Falconer (1960)).

If we have data from k^* replicate populations, the above considerations hold analogously for every replicate $k \in \{1, \ldots, k^*\}$.

We derive variance estimators for the described scenarios in Section S3 of the Supplementary Material (Spitzer, Pelizzola and Futschik (2020)). In Table 3, we present these estimators for $\mathrm{Var}(x_{11})$ and $\mathrm{Var}(x_{21})$ and in Table 4 for $\mathrm{Var}(x_{11k})$ and $\mathrm{Var}(x_{21k})$. To obtain the adapted test statistics for the different scenarios, they can be inserted for $\hat{s}_1^2$ and $\hat{s}_2^2$ in (2), or $\hat{s}_{1k}^2$ and $\hat{s}_{2k}^2$ in (4), respectively. The proposed formulas use estimators $\hat{p}_2$ and $\hat{p}_{2k}$ of $\mathbb{E}[p_2|p_1]$ and $\mathbb{E}[p_{2k}|p_{1k}]$. Also, $\hat{\sigma}^2_{\text{drift}}$ and $\hat{\sigma}^2_{\text{drift-k}}$ are estimators of $\mathrm{Var}(p_2|p_1)$ and $\mathrm{Var}(p_{2k}|p_{1k})$. Choices for these quantities are discussed below.

Notice that different models may apply at different time points. If an experiment involves, for instance, individual sequencing of a sample from the base population and pool sequencing of a sample of the evolved population, the variance estimators should be chosen accordingly: One would take (2) as test statistic with $\hat{s}_1^2$ replaced by $\frac{x_{11}x_{12}}{x_{1+}}$ and $\hat{s}_2^2$ replaced by $r_2(\hat{p}_2(1-\hat{p}_2)(1+\frac{r_2-1}{x_{2+}})+(r_2-1)\frac{x_{2+}-1}{x_{2+}}\hat{\sigma}^2_{\text{drift}})$.

A simple estimator for $\mathbb{E}[p_2|p_1]$ is $\frac{x_{11}}{x_{1+}}$ or $\frac{\hat{x}_{11}}{r_1}$, depending on the number of underlying sampling steps. However, our simulations show that often the distribution of the correspond-

TABLE 3

Estimators $\hat{s}_1^2$ and $\hat{s}_2^2$ of $\mathrm{Var}(x_{11})$ and $\mathrm{Var}(x_{21})$ for different scenarios

	$\hat{s}_1^2$	$\hat{s}_2^2$
1 sampling step*	$x_{1+}\frac{x_{+1}}{n}\frac{x_{+2}}{n}$	$x_{2+}\frac{x_{+1}}{n}\frac{x_{+2}}{n}$
1 sampling step, drift	$\frac{x_{11}x_{12}}{x_{1+}}$	$x_{2+}(\hat{p}_2(1-\hat{p}_2)+(x_{2+}-1)\hat{\sigma}^2_{\text{drift}})$
2 sampling steps	$\frac{\hat{x}_{11}\hat{x}_{12}}{r_1}(1+\frac{r_1-1}{x_{1+}})$	$\frac{\hat{x}_{21}\hat{x}_{22}}{r_2}(1+\frac{r_2-1}{x_{2+}})$
2 sampling steps, drift	$\frac{\hat{x}_{11}\hat{x}_{12}}{r_1}(1+\frac{r_1-1}{x_{1+}})$	$r_2(\hat{p}_2(1-\hat{p}_2)(1+\frac{r_2-1}{x_{2+}})+(r_2-1)\frac{x_{2+}-1}{x_{2+}}\hat{\sigma}^2_{\text{drift}})$

*This is the situation of the classical chi-square test.

TABLE 4
Estimators $\hat{s}^2_{1k}$ and $\hat{s}^2_{2k}$ of $\mathrm{Var}(x_{11k})$ and $\mathrm{Var}(x_{21k})$ for different scenarios

	$\hat{s}^2_{1k}$	$\hat{s}^2_{2k}$
1 sampling step*	$x_{1+k}\frac{x_{+1k}}{n_k}\frac{x_{+2k}}{n_k-1}$	$x_{2+k}\frac{x_{+1k}}{n_k}\frac{x_{+2k}}{n_k-1}$
1 sampling step, drift	$\frac{x_{11k}x_{12k}}{x_{1+k}}$	$x_{2+k}(\hat{p}_{2k}(1-\hat{p}_{2k})+(x_{2+k}-1)\hat{\sigma}^2_{\mathrm{drift-k}})$
2 sampling steps	$\frac{\hat{x}_{11k}\hat{x}_{12k}}{r_{1k}}(1+\frac{r_{1k}-1}{x_{1+k}})$	$\frac{\hat{x}_{21k}\hat{x}_{22k}}{r_{2k}}(1+\frac{r_{2k}-1}{x_{2+k}})$
2 sampling steps, drift	$\frac{\hat{x}_{11k}\hat{x}_{12k}}{r_{1k}}(1+\frac{r_{1k}-1}{x_{1+k}})$	$r_{2k}(\hat{p}_{2k}(1-\hat{p}_{2k})(1+\frac{r_{2k}-1}{x_{2+k}})+(r_{2k}-1)\frac{x_{2+k}-1}{x_{2+k}}\hat{\sigma}^2_{\mathrm{drift-k}})$

*This is the situation of the classical CMH test.

ing p-values is closer to a uniform distribution on [0, 1] when $\mathbb{E}[p_2|p_1]$ is estimated as

$$\frac{\frac{x_{11}}{x_{1+}}+\frac{x_{21}}{x_{2+}}}{2} \quad \text{or} \quad \frac{\frac{\hat{x}_{11}}{r_1}+\frac{\hat{x}_{21}}{r_2}}{2}, \quad \text{respectively.} \tag{6}$$

For a consistent variance estimator of p_2 after t generations of drift, $\hat{\sigma}^2_{\mathrm{drift}}$, we can approximate (5) by

$$\frac{x_{11}x_{12}}{x_{1+}^2}\left(1-\left(1-\frac{1}{2N_e}\right)^t\right) \quad \text{or} \quad \frac{\hat{x}_{11}\hat{x}_{12}}{r_1^2}\left(1-\left(1-\frac{1}{2N_e}\right)^t\right), \quad \text{respectively.} \tag{7}$$

When sequence data for intermediate generations between 0 and t is available, we can use this additional information for the estimation of $\mathbb{E}[p_2|p_1]$ and $\mathrm{Var}(p_2|p_1)$.

Let $t_1 = 0, t_2, \ldots, t_\gamma = t$ be the generations for which sequence data is available, and let $p_1 = f_1, f_2, \ldots, f_\gamma = p_2$ be the corresponding population frequencies of allele 1. Estimating these frequencies in each generation by the corresponding relative sample frequencies $\hat{f}_1, \ldots, \hat{f}_\gamma$, we may proceed as in (6) and use this additional information to estimate $\mathbb{E}[p_2|p_1]$ by

$$\frac{\sum_{i=1}^{\gamma}\hat{f}_i}{\gamma}. \tag{8}$$

Extending (7), the drift variance may also be estimated as

$$\sum_{i=1}^{\gamma-1}\hat{f}_i(1-\hat{f}_i)\left(1-\left(1-\frac{1}{2N_e}\right)^{t_{i+1}-t_i}\right). \tag{9}$$

Analogous estimators may be used for $\mathbb{E}[p_{2k}|p_{1k}]$ and $\mathrm{Var}(p_{2k}|p_{1k})$ in the situation with replicate populations.

As our tests are carried out separately for each locus, they are not directly affected by linkage. It might be worth noting, however, that linkage may lead to significant SNPs which are not the causal targets of selection. This fact is also known as the hitchhiking effect in population genetics and makes it difficult to single out the truly beneficial SNPs. As there is no guarantee that the most significant SNP is the causal target of selection, any test will lead to a set of candidate SNPs containing also several hitchhikers.

When scanning a genomic region or even the full genome, a multiple testing adjustment is required to avoid a large number of false positive SNPs. It is well known that Bonferroni corrections can be very conservative under dependency, and they are therefore not recommended given genetic linkage. Both a Benjamini–Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg (1995)) or the recently proposed harmonic mean p-values

by Wilson (2019) seem reasonable strategies in our context. Notice that the harmonic mean p-values seem to be robust and quite powerful also under positive dependency.

We explore the behavior of the adapted test statistics by computer simulations and present the results in the following section. We focus on the scenario with two sampling steps and genetic drift, since there the additional variance is the largest.

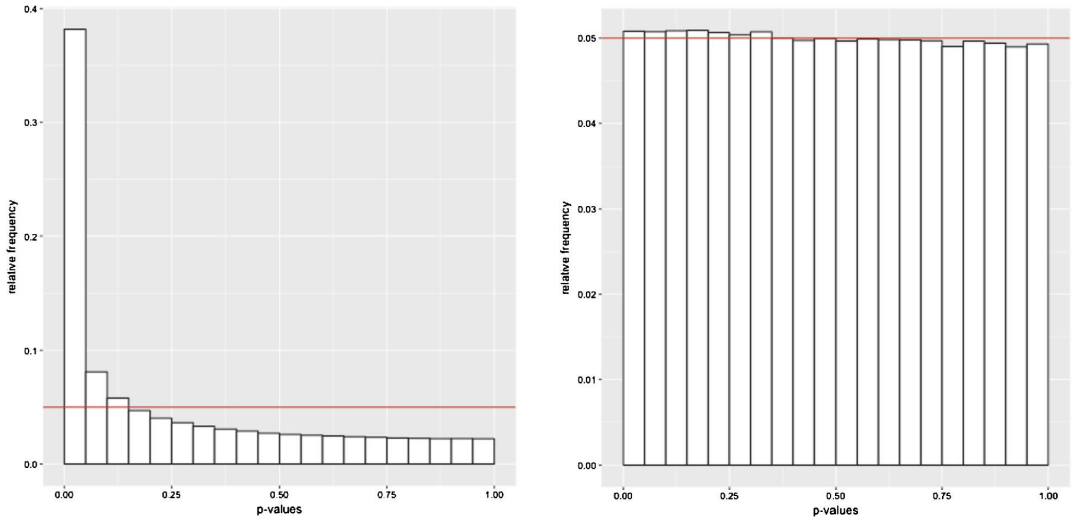
We introduce the following notation: If we do not have data for intermediate generations, we estimate $\mathbb{E}[p_2|p_1]$ and $\mathrm{Var}(p_2|p_1)$, as well as $\mathbb{E}[p_{2k}|p_{1k}]$ and $\mathrm{Var}(p_{2k}|p_{1k})$, by the estimators given in (6) and (7) or their analogs for the kth of k^* replicates. With drift and one sampling step we denote the adapted tests, then $T_{\chi^2}^{\text{1s\&d}}$ and $T_{\text{CMH}}^{\text{1s\&d}}$, with drift and two sampling steps we name them $T_{\chi^2}^{\text{2s\&d}}$ and $T_{\text{CMH}}^{\text{2s\&d}}$. If data for intermediate generations is available, we apply the estimators (8) and (9) or their analogs for the kth replicate. In the case of drift, we denote the adapted tests by $T_{\chi^2}^{\text{1s\&d-ig}}$ and $T_{\text{CMH}}^{\text{1s\&d-ig}}$ (one sampling step), and $T_{\chi^2}^{\text{2s\&d-ig}}$ and $T_{\text{CMH}}^{\text{2s\&d-ig}}$ (two sampling steps).

3. Simulation results. We carried out extensive simulations in *R* (R-Core-Team (2018)) in order to explore the behavior of the adapted tests described in the previous section. We simulated genetic drift using the package *poolSeq* (Taus, Futschik and Schlötterer (2017)). When we encountered loci with frequency 0 for one allele in the base population but a positive frequency in a later generation, we changed the allele count from 0 to 1 in order to always obtain a well-defined test statistic. When these situations occur with real data, there are different possible explanations: Either a mutation arose in a later generation and the allele really was not present before, or the frequency of the respective allele is low, but not 0, in the base population and the allele was just by chance not sampled or amplified in the sequencing process. Since mutation rates are usually low over such a time span (Burke et al. (2010)), the latter scenario is the more likely one. Finally, if the frequency in the later generation is very low, the nonzero frequency may also be due to a sequencing error. Overall, our method to deal with this phenomenon seems to be a pragmatic compromise.

We first provide results under the null hypothesis and then examine the power of our adapted test statistics. After that we compare our adapted tests to other state-of-the-art methods. We first simulate allele frequencies in generation 0 uniformly distributed on $[0, 1]$ to give all possible true allele frequencies the same weight. At the end of the section, an allele frequency distribution for generation 0 is used that resembles the one encountered in experimental data available to us.

Null distribution. In our simulations we set $N_e = 300$ and used 1000 as sample size of alleles that were sequenced at generations 0 and 60; the sequencing coverage was chosen Poisson distributed with mean 80. These parameter choices were motivated by the real data for *Drosophila* taken from Barghi et al. (2017) and discussed in Section 4.

To control the type I error, it is desirable that the p-values of a test are uniformly distributed on $[0, 1]$ or, at least, stochastically larger than uniform if the null hypothesis is true. Indeed, under the neutral Wright–Fisher model we observe that the distribution of the p-values belonging to the adapted tests is close to a uniform distribution and that the tests control the 5% significance level. In contrast, the nonadapted tests show a huge excess of small p-values. Based on 10^6 simulated loci, Figure 2 displays the distribution of the p-values for the classical CMH test with three replicates and the test adapted to drift, sampling and pool sequencing $T_{\text{CMH}}^{\text{2s\&d}}$. For the adapted chi-square test $T_{\chi^2}^{\text{2s\&d}}$, the distribution of the p-values is slightly further from a uniform distribution but controls the 5% significance threshold (Supplementary Material (Spitzer, Pelizzola and Futschik (2020)), Figure S1). We obtain similar results in situations with drift and only one sampling step (Supplementary Material



(a) T_{CMH} (b) $T_{\mathrm{CMH}}^{2\mathrm{s\&d}}$

FIG. 2. *Distribution of the p-values obtained from* T_{CMH} (*a*) *and* $T_{\mathrm{CMH}}^{2\mathrm{s\&d}}$ (*b*). *Horizontal line indicates frequency of* 5%. *Simulation setup*: 10^6 *neutral loci with true allele frequencies in base population uniformly distributed on* $[0, 1]$, $N_e = 300$, *allele sample size* 1000, *pool sequencing with Poisson distributed coverage* ($\mu = 80$), *sequence data for generations* 0 *and* 60 *and three replicate populations*.

(Spitzer, Pelizzola and Futschik (2020)), Figure S2). With two underlying sampling steps but without drift, a situation occurring, for example, in GWAS using pool sequencing data (e.g., Bastide et al. (2013), Endler et al. (2016)), the adapted tests show again an improved performance (Supplementary Material (Spitzer, Pelizzola and Futschik (2020)), Figure S3).

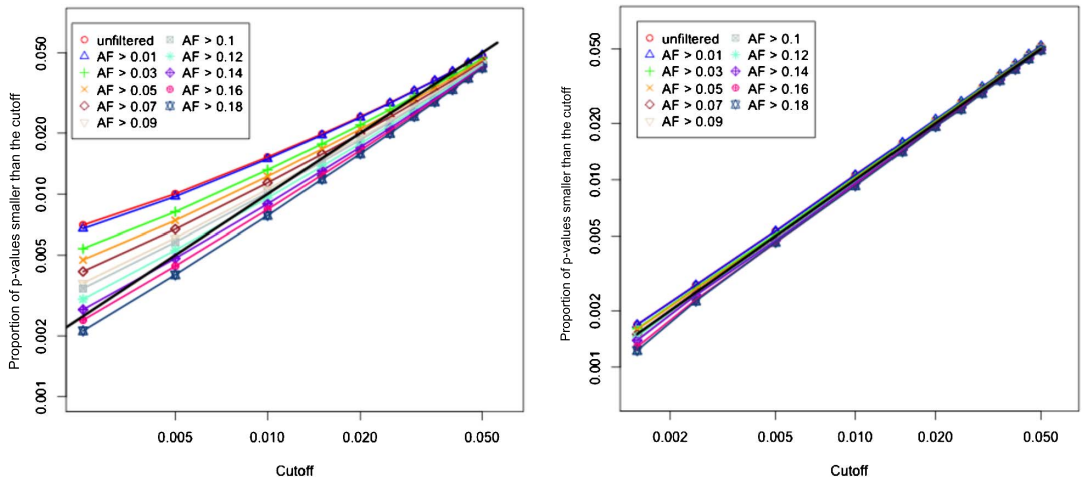
In genome-wide applications, corrections for multiple testing are usually needed, and significance levels much below 0.05 become important. To check the type I error control in such a situation, we did a simulation analysis for 10^6 loci based on sequence data with the parameter choices as above (but five replicate populations in case of the CMH test). As shown in Figure 3, especially $T_{\chi^2}^{2\mathrm{s\&d}}$ turns out to be anticonservative for very small significance levels.

To understand this issue, it should be noted that the distributional approximations involved in these tests are less reliable for loci with a very small or very high allele frequency.

An obvious remedy would be to introduce a threshold value ζ and only consider loci with frequency of the minor (less frequent) allele larger than ζ in the base population. According to Figure 3, filtering has the desired effect, if ζ is chosen large enough.

One disadvantage of this approach is that by filtering out SNPs with small and large allele frequencies, we exclude a lot of potentially selected loci. We therefore explore also other approaches to resolve the issue: If sequence data is available, not only from two time points but also from intermediate generations, we can modify the adapted tests by taking the additional information into account, resulting in the test statistics $T_{\chi^2}^{1\mathrm{s\&d\text{-}ig}}$, $T_{\chi^2}^{2\mathrm{s\&d\text{-}ig}}$, $T_{\mathrm{CMH}}^{1\mathrm{s\&d\text{-}ig}}$ and $T_{\mathrm{CMH}}^{2\mathrm{s\&d\text{-}ig}}$; see Section 2.1. Simulations with the same parameters as in Figure 3 but, additionally, with sequence data every 10 generations, show that $T_{\chi^2}^{2\mathrm{s\&d\text{-}ig}}$ and $T_{\mathrm{CMH}}^{2\mathrm{s\&d\text{-}ig}}$ hold the 5% level or are conservative for all significance levels without filtering out small and large allele frequencies (Supplementary Material (Spitzer, Pelizzola and Futschik (2020)), Figure S4).

If time series data is not available, a p-value correction may be applied by fitting the distribution F_p of p-values simulated under the null hypothesis and transforming the p-values to uniform using $F_p(\cdot)$. In Section S4 of the Supplementary Material (Spitzer, Pelizzola and Futschik (2020)), we propose a parametric choice of F_p for correcting small (potentially sig-



(a) $T_{\chi^2}^{2s\&d}$ (b) $T_{\mathrm{CMH}}^{2s\&d}$

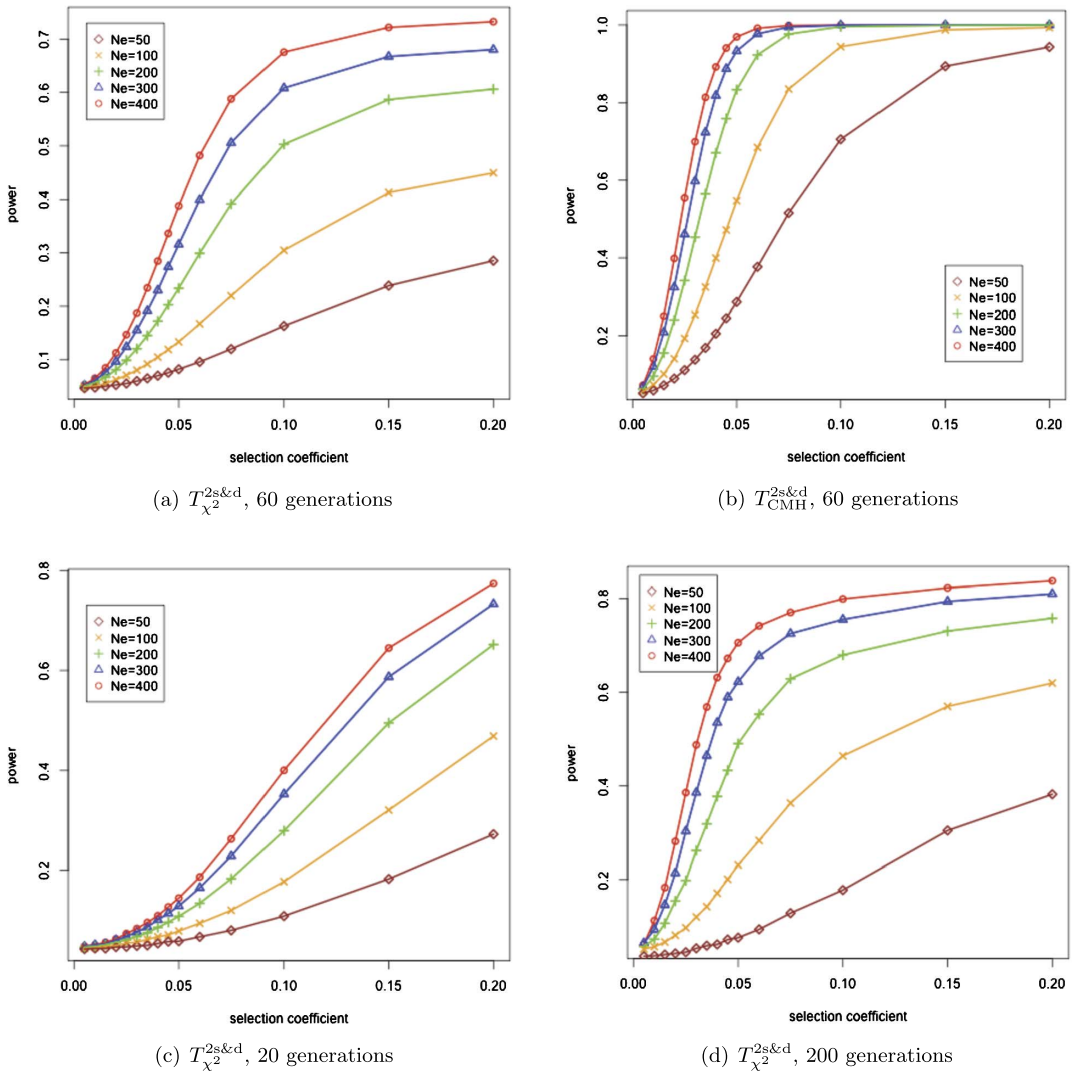
FIG. 3. *Proportion of loci with p-value smaller than cutoff value against cutoff value for different minimum values of the allele frequencies of both alleles in the base population, in* (*a*) *for* $T_{\chi^2}^{2s\&d}$ *and in* (*b*) *for* $T_{\mathrm{CMH}}^{2s\&d}$. *The black solid line is the angle bisector. Simulation setup*: 10^6 *neutral loci with true allele frequencies in base population uniformly distributed on* $[0, 1]$, $N_e = 300$, *allele sample size* 1000, *pool sequencing with Poisson distributed coverage* ($\mu = 80$), *sequence data for generations* 0 *and* 60 *and five replicate populations in* (*b*).

nificant) p-values and show that it leads to a substantial improvement of the null distribution of p-values obtained with $T_{\chi^2}^{2s\&d}$.

Power. We additionally carried out simulations involving 10^5 selected loci in order to examine the power of the adapted tests at a significance threshold of $\alpha = 0.05$. We first consider a realistic set of standard parameter choices: $N_e = 300$, sample size 1000, coverage 100, and sequence data available for generations 0 and 60. To also investigate the influence of these model parameters, we considered a set of alternative values for each of them.

Not surprisingly, the power of $T_{\mathrm{CMH}}^{2s\&d}$ is higher than the power of $T_{\chi^2}^{2s\&d}$ because the amount of information increases with more replicates. As also expected, the power increases with the selection coefficient s and the effective population size N_e. Figure 4(a) shows the power of $T_{\chi^2}^{2s\&d}$ for different values of s and N_e. Figure 4(b) displays the same for $T_{\mathrm{CMH}}^{2s\&d}$ with five replicates. Comparing Figures 4(c) and 4(d), which are for the same scenario as in Figure 4(a) but with 20 and 200 generations of evolution, we see that for small selection coefficients the power increases with the number of generations. This is since more generations of evolution lead to larger frequency differences between base and evolved populations unless the selection coefficients are large. Under strong selection many alleles soon reach a frequency close to 1 (fixation), and, hence, the signal of selection does not become stronger anymore with more generations. On the other hand, the drift variance that we calculate for the denominator of the test statistic increases with every generation which reduces the power. The gain in power due to more generations of evolution and the loss in power due to fixation may cancel each other out, leading to the plateauing effect at around 0.8 we observe in Figure 4(d). The impact of the number of generations on the power is qualitatively the same for $T_{\mathrm{CMH}}^{2s\&d}$. (Supplementary Material (Spitzer, Pelizzola and Futschik (2020)), Figure S5).

The influence of sample size and coverage is shown in Figure 5, in (a) for $T_{\chi^2}^{2s\&d}$ and in (b) for $T_{\mathrm{CMH}}^{2s\&d}$. The power increases with the sample size and with the coverage. Since the coverage values are an order of magnitude smaller than the values for the sample size, the effect is much more pronounced for the coverage.



(a) $T_{\chi^2}^{2s\&d}$, 60 generations

(b) $T_{\mathrm{CMH}}^{2s\&d}$, 60 generations

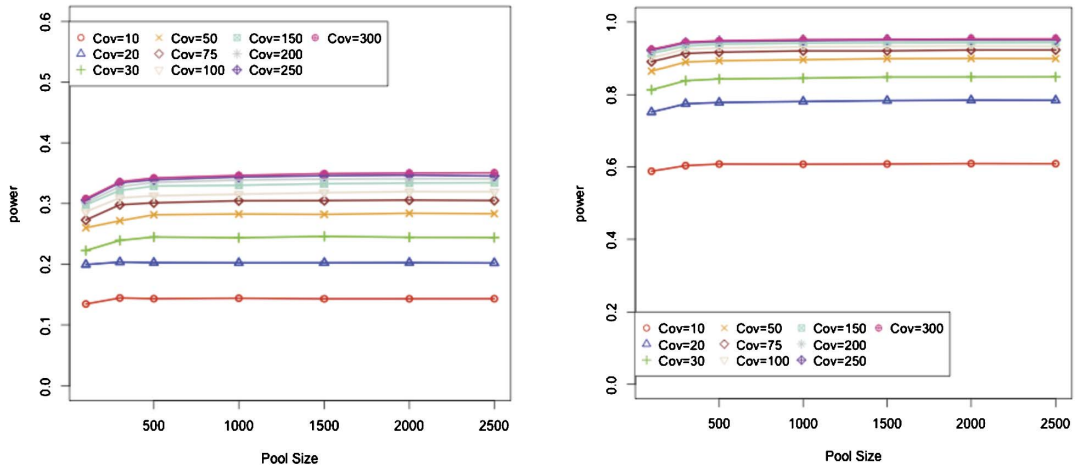
(c) $T_{\chi^2}^{2s\&d}$, 20 generations

(d) $T_{\chi^2}^{2s\&d}$, 200 generations

FIG. 4. *Power vs. selection coefficients for different values of the effective population size N_e, in* (*a*), (*c*) *and* (*d*) *for* $T_{\chi^2}^{2s\&d}$ *and in* (*b*) *for* $T_{\mathrm{CMH}}^{2s\&d}$. *Simulation setup*: 10^5 *loci for each selection coefficient with true allele frequencies in base population uniformly distributed on* [0, 1], *allele sample size* 1000, *pool sequencing with coverage* 100 *and five replicate populations in* (*b*); (*a*) *and* (*b*) *sequence data generations* 0 *and* 60, (*c*) *sequence data generations* 0 *and* 20, (*d*) *sequence data generations* 0 *and* 200.

The power of $T_{\chi^2}^{2s\&d\text{-ig}}$ and $T_{\mathrm{CMH}}^{2s\&d\text{-ig}}$ is similar to the power of $T_{\chi^2}^{2s\&d}$ and $T_{\mathrm{CMH}}^{2s\&d}$; see Table 5. In general, $T_{\chi^2}^{2s\&d\text{-ig}}$ and $T_{\mathrm{CMH}}^{2s\&d\text{-ig}}$ are affected in the same way by the parameter choices as $T_{\chi^2}^{2s\&d}$ and $T_{\mathrm{CMH}}^{2s\&d}$ (not shown).

Method comparison. We compared the performance of our adapted test statistics to other state-of-the-art methods. As summaries we considered the type I error, the power and the run time. We looked at the classical chi-square test with the modified rejection cutoff (OrozcoterWengel et al. (2012)) and the LLS approach for detecting selection of Taus, Futschik and Schlötterer (2017). As mentioned in the Introduction, the LLS method is faster than other methods such as CLEAR (Iranmehr et al. (2017)). Still, the computation times encountered with the LLS method are high, and we restricted our simulations, therefore, here to 10^4 loci



(a) $T_{\chi^2}^{2s\&d}$ (b) $T_{\mathrm{CMH}}^{2s\&d}$

FIG. 5. *Power vs. sample size of alleles (pool size) for different pool sequencing coverages, in (a) for* $T_{\chi^2}^{2s\&d}$, *and in (b) for* $T_{\mathrm{CMH}}^{2s\&d}$. *Simulation setup*: 10^5 *loci for each value of the pool size with true allele frequencies in base population uniformly distributed on* [0, 1], $N_e = 300$, *selection coefficient* 0.05, *sequence data for generations* 0 *and* 60, *and five replicate populations in (b).*

of which 10% are under selection. The other parameters were chosen again $N_e = 300$, allele sample size 1000, coverage Poisson distributed with mean 80 and 60 generations of drift.

Compared with the LLS method by Taus, Futschik and Schlötterer (2017), $T_{\chi^2}^{2s\&d}$ and $T_{\chi^2}^{2s\&d\text{-ig}}$ have a similar power to detect selection while being much faster; see Table 5. The chi-square test with the modified rejection cutoff (Orozco-terWengel et al. (2012)) needs a similar computation time as our tests but has much lower power. Indeed, the simulated rejection boundary makes it over conservative.

TABLE 5

Comparison of type I error, power and running times in seconds for different tests of selection. Simulation setup: 10^4 *loci with true allele frequencies in base population uniformly distributed on* [0, 1], 10% *of the loci under selection, selection coefficients exponentially distributed with mean* 0.1, $N_e = 300$, *allele sample size* 1000, *pool sequencing with Poisson distributed coverage* ($\mu = 80$), *sequence data for generations* 0 *and* 60, *additionally for generations* 10, 20, 30, 40, 50 *when* $T_{\chi^2}^{2s\&d\text{-ig}}$, $T_{\mathrm{CMH}}^{2s\&d\text{-ig}}$ *and LLS are applied; five replicate populations with* $T_{\mathrm{CMH}}^{2s\&d}$, *and* $T_{\mathrm{CMH}}^{2s\&d\text{-ig}}$

Test	Type I Error	Power	Time
$T_{\chi^2}^{2s\&d}$	0.050	0.417	0.005
$T_{\chi^2}^{2s\&d\text{-ig}}$	0.046	0.490	0.369
$T_{\mathrm{CMH}}^{2s\&d}$ with five replicates	0.050	0.761	0.014
$T_{\mathrm{CMH}}^{2s\&d\text{-ig}}$ with five replicates	0.049	0.756	0.487
T_{χ^2}	0.374	0.751	0.144
T_{χ^2} with modified rejection cutoff (Orozco-terWengel et al. (2012))	0.00034	0.226	0.144
LLS* (Taus, Futschik and Schlötterer (2017))	0.047	0.456	33,568.010

*Assume diploids, dominance is set to 0.5, the method to estimate selection is set to "LLS" and p-values are simulated with N.pval set to 1000, which means that 1000 simulations are performed to estimate the p-values.

TABLE 6

Proportion of positively detected SNPs for each tool. Simulation setup: 10^4 *SNPs under selection, selection coefficients exponentially distributed with mean* 0.06, *and* 10^4 *neutrally evolving SNPs are simulated with* $N_e = 300$, *allele sample size* 1000, *pool sequencing with Poisson distributed coverage* ($\mu = 80$), *sequence data for generations* 0 *and* 60, *additionally, for generations* 10, 20, 30, 40, 50 *when* $T_{\text{CMH}}^{\text{2s\&d-ig}}$ *and LLS are applied, three replicate populations*

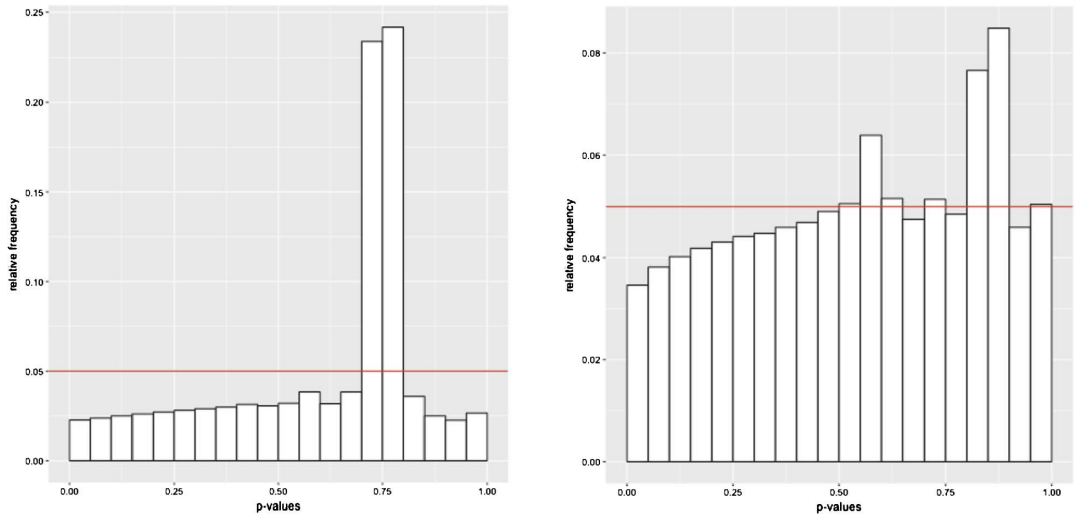
	Truth	
	Positive	Negative
Positive for $T_{\text{CMH}}^{\text{2s\&d-ig}}$	0.5926	0.0472
Positive for $T_{\text{CMH}}^{\text{2s\&d}}$	0.5485	0.0479
Positive for T_{CMH}	0.7708	0.3448
Positive for T_{CMH} with modified rejection cutoff (Orozco-terWengel et al. (2012))	0.0445	0.003
Positive for LLS	0.5663	0.0478

We also ran simulations with slightly different parameter choices and computed the proportion of true and false positives. For the results and details on the parameter values, see Table 6. The results, as well as the ones in Table 5, illustrate the favourable performance of our methods.

Our proposed test statistics $T_{\chi^2}^{\text{1s\&d-ig}}$ and $T_{\text{CMH}}^{\text{1s\&d-ig}}$ have also been compared to several other tests in an extensive benchmarking work in Vlachos et al. (2019) for E&R studies. The comparison considers three realistic selection scenarios. It turns out that our proposed test statistics rank among the top methods for each scenario, both in terms of power and time.

Alternative starting allele frequency distribution. In the previous simulations, the true allele frequencies in the base populations were chosen uniformly distributed on [0, 1]. In general, uniformly distributed allele frequencies are not common in natural populations. Therefore, we also looked for a more realistic distribution of allele frequencies. As an example we consider the U-shaped beta distribution proposed in Jónás et al. (2016). The U-shape depicts the observed excess of high and low allele frequencies in a folded site frequency spectrum where it is not known which allele is ancestral and which one is derived. Following this example, we simulated the true allele frequencies in generation 0 from a beta distribution with parameters 0.2 and 0.2. The other parameters remained unchanged compared to the previous simulations. As the chosen distribution produces a higher proportion of allele frequencies close to the boundaries, the distribution of the p-values becomes less uniform (see Figure 6). The spikes are caused by discreteness phenomena occurring for very low starting allele frequencies. However, $T_{\chi^2}^{\text{2s\&d-ig}}$ and $T_{\text{CMH}}^{\text{2s\&d-ig}}$ turned out to be quite conservative (even for significance levels smaller than 0.05) when sequence data every 10 generations is available. At the same time, power values of the adapted tests are lower than when under uniformly distributed starting allele frequencies (Supplementary Material (Spitzer, Pelizzola and Futschik (2020)), Figure S6). An explanation would be that even SNPs with positively selected alleles have a large probability of being lost due to drift in early generations, if the initial allele frequency is very low. For such SNPs we do not have any power of detection.

In particular, we calculated a power of approximately 0.59 for $T_{\text{CMH}}^{\text{2s\&d}}$ based on 10^5 simulations with selection coefficient 0.1, allele sample size 1000, pool sequencing in generations 0 and 60 with coverage Poisson distributed with mean 80 and five replicates. When additional sequence data for generations 10, 20, 30, 40, and 50 was simulated, we obtained a power of approximately 0.60 using $T_{\text{CMH}}^{\text{2s\&d-ig}}$ as test statistic.



(a) $T_{\chi^2}^{2s\&d\text{-}ig}$ (b) $T_{CMH}^{2s\&d\text{-}ig}$

FIG. 6. *Distribution of the p-values obtained from* $T_{\chi^2}^{2s\&d\text{-}ig}$ (*a*) *and* $T_{CMH}^{2s\&d\text{-}ig}$ (*b*). *Horizontal line indicates frequency of* 5%. *Simulation setup*: 10^6 *neutral loci with true allele frequencies in base population beta distributed with parameters* 0.2 *and* 0.2, $N_e = 300$, *allele sample size* 1000, *pool sequencing with Poisson distributed coverage* ($\mu = 80$), *sequence data every* 10 *generations from generation* 0 *to* 60 *and three replicate populations in* (*b*).

4. Application to experimental data from *Drosophila*. Here, we consider data from an E&R experiment on *Drosophila simulans* as described in Barghi et al. (2017). In this publication the classical CMH test has been used to infer candidates of selection. Allele frequency measurements were taken from three replicate populations at generations 0 and 60. All flies were maintained under a cycling routine to stimulate temperature adaptation (12 hours at 18°C and 12 hours at 28°C with light for the day).

In the original paper neutral simulations have been used to define a cutoff that leads to 2% false positive SNPs under the simulated global null model. This cutoff has then been taken as a threshold for the p-values obtained with the CMH test applied to the real data; see OrozcoterWengel et al. (2012) for a more detailed description. Assuming that the null model applies to most of the SNPs, this approach will lead to approximately 80,000 false positive SNPs and an unclear false discovery rate given a genome of approximately four million SNPs as considered in Barghi et al. (2017).

An advantage of our approach is that the resulting proper p-values can be combined with a standard procedure such as harmonic mean p-values by Wilson (2019) or Benjamini–Hochberg (Benjamini and Hochberg (1995)) that control for multiple testing. Another advantage of our method is that it will lead to a more proper ranking of the p-values, as the amount of error incurred with the classical test statistics depends on the relative magnitudes of the variance components (drift variance, sequencing coverage and sample size) and will therefore vary between SNPs.

Since the DNA from the whole population was used in the pool sequencing step, we applied $T_{\chi^2}^{1s\&d}$ and $T_{CMH}^{1s\&d}$ as test statistics using the model parameters specified in Barghi et al. (2017). We only present the results for the modified CMH test on the entire data set. A Manhattan plot of the SNP positions versus the logarithm (base 10) of the p-values corrected for multiple testing with the Benjamini–Hochberg method (Benjamini and Hochberg (1995)) is shown in Figure 7. The computation time needed for this analysis was only about 20 seconds on a standard laptop. We infer more significant SNPs than Barghi et al. (2017) (0.0049% and

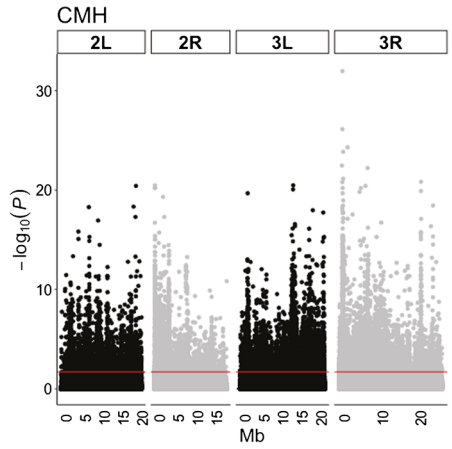


FIG. 7. *Manhattan plot for* $T_{\text{CMH}}^{\text{1s\&d}}$ *applied to real data from Drosophila simulans from Barghi et al.* (2017).

0.0002% of the total number of tested SNPs, respectively). This result is concordant with our simulations showing that our method has higher power compared to the modified rejection cutoff (Orozco-terWengel et al. (2012)) which is applied in Barghi et al. (2017).

We also applied LLS, $T_{\text{CMH}}^{\text{1s\&d}}$ and T_{CMH} (with the modified rejection cutoff) to a small region (5000 SNPs) of chromosome 2L from the data in Barghi et al. (2017). Here, after correcting for multiple testing with the Benjamini–Hochberg procedure (Benjamini and Hochberg (1995)), we focused on 30 SNPs shown in the Manhattan plot (Supplementary Material (Spitzer, Pelizzola and Futschik (2020)), Figure S7). As can be seen, the order of the SNPs differs considerably between $T_{\text{CMH}}^{\text{1s\&d}}$ and T_{CMH}. This can be explained by the differences in coverage values between SNPs which correctly contribute to the variance only with our method.

5. Discussion. With population genetic applications in mind, we propose modified test statistics for the chi-square and the CMH tests in scenarios with overdispersion, that is, more variance in the data than considered by the original tests. Compared with the classical versions of these tests that are still commonly applied, our approach does not require simulations to find a cutoff for the test statistic. Our proposed approach can also be used instead of Fisher's exact test which faces the same problems in this context. Our inclusion of proper variance terms leads to a better performance both with respect to power and type I error. Using the classical tests does not even lead to a proper ranking of the SNPs as the amount of overdispersion varies between the SNPs. While more sophisticated testing procedures have also been proposed, they are usually considerably more time consuming, especially when applied on a genome-wide scale.

While overdispersion is also known in other applications (such as complex surveys), our underlying model requires a different adaptation of the test statistics. In a first step, we therefore expressed the test statistics in dependence of the variances of entries in the contingency tables. Using this general form, suitable tests can be obtained in any situation with overdispersion provided that the required variances can be properly estimated. As a special case, we then derived explicit formulas for the adapted test statistics for use in different types of E&R experiments under scenarios including all or some of the components of variance: actual sampling, pool sequencing and genetic drift.

Our test statistics do not only provide a more appropriate error control but have also considerably larger power than the classical tests combined with computer simulations such as proposed in Orozco-terWengel et al. (2012).

As the power of our tests increases with the sequencing coverage, N_e, and allele sample size, experimenters could use our results on power to set up their experimental design. For more detailed recommendations on experimental design, see, for example, Kofler and Schlötterer (2014).

Compared with more sophisticated methods, our adapted tests have approximately the same power but need much less computation time, which is an important factor when the whole genome is scanned for traces of selection. Our tests are, for example, 10^5 times faster than the LLS method by Taus, Futschik and Schlötterer (2017).

When sequence data is available for two time points only, our tests are not always conservative for very small allele frequencies in the base population. This problem is related to inaccuracies of the normal approximation of proportions close to zero and one. In this case, we propose a correction of the p-values based on simulations under the null model. Notice, however, that this additional correction is usually not needed when data is available also at intermediate time points, as supported by our simulated scenarios. Therefore we recommend experimenters to sequence intermediate generations, if the budget allows for it.

Besides E&R experiments, the presented tests are also applicable in other situations. For instance, with GWAS using data from pool sequencing there are two sampling steps but no drift. When testing for differences between traits with the chi-square or the CMH test, the variance estimators presented in line 3 of Tables 3 and 4 should therefore be used.

Another application would be tests for population specific adaptation. A typical scenario in this context is the split of one population into two separate populations which are then kept under different conditions for a number of generations. Our tests can be used to check whether differences in allele frequencies between the two populations go beyond what is expected by drift and sampling noise. The modifications needed for $T_{\chi^2}^{1s\&d}$ are explained in the Appendix. The other tests can be dealt with analogously. For an application, see Tobler, Hermisson and Schlötterer (2015) where population specific adaptation is considered for *Drosophila* in a hot and a cold environment in order to study thermal adaptation.

With appropriate estimates of the drift variance, our tests can also be used for the comparison of allele frequencies in natural (sub)populations in order to detect loci under selection. The latter scenario is, for example, considered by Beaumont and Balding (2004) as well as Foll and Gaggiotti (2008) that have developed hierarchical Bayesian models for the situation.

We implemented the adapted tests in an *R* package called ACER which can be downloaded at https://github.com/MartaPelizzola/ACER. The source code is additionally found in the Supplementary Material to this paper (Spitzer, Pelizzola and Futschik (2020)).

Our results suggest that our adapted test statistics provide fast, reliable and powerful methods to detect selection. Hence, they have the potential to considerably facilitate the inference of selected loci in population genetics, in particular in the context of E&R.

APPENDIX: TESTING FOR POPULATION SPECIFIC ADAPTATION WITH THE CHI-SQUARE TEST

Here, capital letters indicate random variables (except from N_e).

Suppose population 1 and population 2 stem both from the same base population. Assume that population 1 has evolved for t_1 generations and population 2 has evolved for t_2 generations. We assume that one sampling step was involved to obtain allele frequency counts. The adapted chi-square test (2) may be used with the allele frequencies from the evolved populations 1 and 2. In the following, we state formulas for the required variance estimators $\hat{s}_1^2$ and $\hat{s}_2^2$.

To obtain drift variance estimates, we assume that the base population has also been sequenced. Let p_0 be the allele frequency of allele 1 in the base population, X_{01} the corresponding counts of allele 1 and x_{0+} the total number of allele counts from the base population.

Furthermore, $\hat{p}_1$ and $\hat{p}_2$ are estimators of $\mathbb{E}[P_1|p_0]$ and $\mathbb{E}[P_2|p_0]$, respectively, also $\hat{\sigma}^2_{\text{drift1}}$ and $\hat{\sigma}^2_{\text{drift2}}$ are estimators of $\text{Var}(P_1|p_0)$ and $\text{Var}(P_2|p_0)$, respectively.

Analogous to the variance estimator for the evolved population in line 2 of Table 3 (but with a different notation), we can estimate $\hat{s}_1^2$ and $\hat{s}_2^2$ as

$$\hat{s}_1^2 = x_{1+}\big(\hat{p}_1(1-\hat{p}_1) + (x_{1+}-1)\hat{\sigma}^2_{\text{drift1}}\big), \tag{A1}$$

and

$$\hat{s}_2^2 = x_{2+}\big(\hat{p}_2(1-\hat{p}_2) + (x_{2+}-1)\hat{\sigma}^2_{\text{drift2}}\big). \tag{A2}$$

Analogous to equations (6) and (7), we can set the estimators for $\hat{p}_1$ and $\hat{p}_2$ as

$$\hat{p}_1 = \frac{\frac{X_{01}}{x_{0+}} + \frac{X_{11}}{x_{1+}}}{2}, \tag{A3}$$

$$\hat{p}_2 = \frac{\frac{X_{01}}{x_{0+}} + \frac{X_{21}}{x_{2+}}}{2}, \tag{A4}$$

and the respective estimators for the variance as

$$\hat{\sigma}^2_{\text{drift1}} = \frac{X_{01}(x_{0+}-X_{01})}{x_{0+}^2}\left(1-\left(1-\frac{1}{2N_e}\right)^{t_1}\right), \tag{A5}$$

$$\hat{\sigma}^2_{\text{drift2}} = \frac{X_{01}(x_{0+}-X_{01})}{x_{0+}^2}\left(1-\left(1-\frac{1}{2N_e}\right)^{t_2}\right). \tag{A6}$$

Acknowledgments. Kerstin Spitzer previously was Kerstin Gärtner.

We thank Neda Barghi and Christian Schlötterer for providing the data and helpful comments, as well as Thomas Taus for his input on the acceleration of the R-code. We are also grateful to the Editor and the reviewers for their valuable comments. This work has been supported by the Austrian Science Fund (FWF Doctoral Program "Vienna Graduate School of Population Genetics", DK W1225-B20). The Vienna Graduate School of Population Genetics is hosted at the Institut für Populationsgenetik, Vetmeduni Vienna, 1210 Vienna, Austria.

SUPPLEMENTARY MATERIAL

Supplement A: Calculations and additional figures (DOI: 10.1214/19-AOAS1301SUPPA; .pdf). Calculations of the formulas presented in this article and additional figures.

Supplement B: Code for adapted chi-square test (DOI: 10.1214/19-AOAS1301SUPPB; .zip). Code for the adapted chi-square test implemented in the R-package ACER, downloadable at https://github.com/MartaPelizzola/ACER.

Supplement C: Code for adapted CMH test (DOI: 10.1214/19-AOAS1301SUPPC; .zip). Code for the adapted CMH test implemented in the R-package ACER, downloadable at https://github.com/MartaPelizzola/ACER.

REFERENCES

AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, New York. MR1914507 https://doi.org/10.1002/0471249688

BARGHI, N., TOBLER, R., NOLTE, V. and SCHLÖTTERER, C. (2017). *Drosophila simulans*: A species with improved resolution in evolve and resequence studies. *G3*: *Genes*, *Genomes*, *Genetics* **7** 2337–2343. https://doi.org/10.1534/g3.117.043349

BASTIDE, H., BETANCOURT, A., NOLTE, V., TOBLER, R., STÖBE, P., FUTSCHIK, A. and SCHLÖTTERER, C. (2013). A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. *PLoS Genet*. **9** e1003534. https://doi.org/10.1371/journal.pgen.1003534.

BEAUMONT, M. A. and BALDING, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* **13** 969–980. https://doi.org/10.1111/j.1365-294X.2004.02125.x.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BOLLBACK, J. P., YORK, T. L. and NIELSEN, R. (2008). Estimation of 2Nes from temporal allele frequency data. *Genetics* **179** 497–502. https://doi.org/10.1534/genetics.107.085019

BURKE, M. K., DUNHAM, J. P., SHAHRESTANI, P., THORNTON, K. R., ROSE, M. R. and LONG, A. D. (2010). Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* **467** 587–590. https://doi.org/10.1038/nature09352.

ENDLER, L., BETANCOURT, A. J., NOLTE, V. and SCHLÖTTERER, C. (2016). Reconciling differences in poolgwas between populations: A case study of female abdominal pigmentation in *Drosophila melanogaster*. *Genetics* **202** 843–855. https://doi.org/10.1534/genetics.115.183376.

EWENS, W. J. (2004). *Mathematical Population Genetics. I: Theoretical Introduction*, 2nd ed. *Interdisciplinary Applied Mathematics* **27**. Springer, New York. MR2026891 https://doi.org/10.1007/978-0-387-21822-9

FALCONER, D. S. (1960). *Introduction to Quantitative Genetics*. The Ronald Press Company, New York.

FEDER, A. F., KRYAZHIMSKIY, S. and PLOTKIN, J. B. (2014). Identifying signatures of selection in genetic time series. *Genetics* **196** 509–522. https://doi.org/10.1534/genetics.113.158220.

FOLL, M. and GAGGIOTTI, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics* **180** 977–993. https://doi.org/10.1534/genetics.108.092221.

FOLL, M., SHIM, H. and JENSEN, J. D. (2015). WFABC: A Wright–Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Mol. Ecol.* **15** 87–98. https://doi.org/10.1111/1755-0998.12280.

GRIFFIN, P. C., HANGARTNER, S. B., FOURNIER-LEVEL, A. and HOFFMANN, A. A. (2017). Genomic trajectories to desiccation resistance: Convergence and divergence among replicate selected *Drosophila* lines. *Genetics* **205** 871–890. https://doi.org/10.1534/genetics.116.187104

ILLINGWORTH, C. J. R., PARTS, L., SCHIFFELS, S., LITI, G. and MUSTONEN, V. (2012). Quantifying selection acting on a complex trait using allele frequency time series data. *Mol. Biol. Evol.* **29** 1187–1197. https://doi.org/10.1093/molbev/msr289.

IRANMEHR, A., AKBARI, A., SCHLÖTTERER, C. and BAFNA, V. (2017). CLEAR: Composition of Likelihoods for Evolve and Resequence Experiments. *Genetics* 1011–1023. https://doi.org/10.1101/080085.

JÓNÁS, Á., TAUS, T., KOSIOL, C., SCHLÖTTERER, C. and FUTSCHIK, A. (2016). Estimating the effective population size from temporal allele frequency changes in experimental evolution. *Genetics* **204** 723–735. https://doi.org/10.1534/genetics.116.191197

KOFLER, R., PANDEY, R. V. and SCHLÖTTERER, C. (2011). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27** 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

KOFLER, R. and SCHLÖTTERER, C. (2014). A guide for the design of evolve and resequencing studies. *Mol. Biol. Evol.* **31** 474–483. https://doi.org/10.1093/molbev/mst221.

LEVY, S. F., BLUNDELL, J. R., VENKATARAM, S., PETROV, D. A., FISHER, D. S. and SHERLOCK, G. (2015). Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519** 181–186. https://doi.org/10.1038/nature14279.

LOHR, S. L. (2010). *Sampling: Design and Analysis*, 2nd ed. Brooks/Cole Cengage Learning, Boston, MA. MR3057878

MALASPINAS, A. S., MALASPINAS, O., EVANS, S. N. and SLATKIN, M. (2012). Estimating allele age and selection coefficient from time-serial data. *Genetics* **192** 599–607. https://doi.org/10.1534/genetics.112.140939.

MATHIESON, I. and MCVEAN, G. (2013). Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193** 973–984. https://doi.org/10.1534/genetics.112.147611.

MCDONALD, J. H. (2014). *Handbook of Biological Statistics*, 3rd ed. Sparky House Publishing, Baltimore, MD.

NOUHAUD, P., TOBLER, R., NOLTE, V. and SCHLÖTTERER, C. (2016). Ancestral population reconstitution from isofemale lines as a tool for experimental evolution. *Ecol. Evol.* **6** 7169–7175. https://doi.org/10.1002/ece3.2402.

OROZCO-TERWENGEL, P., KAPUN, M., NOLTE, V., KOFLER, R., FLATT, T. and SCHLÖTTERER, C. (2012). Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Mol. Ecol.* **21** 4931–4941. https://doi.org/10.1111/j.1365-294X.2012.05673.x.

R-CORE-TEAM (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

REMOLINA, S. C., CHANG, P. L., LEIPS, J., NUZHDIN, S. V. and HUGHES, K. A. (2012). Genomic basis of aging and life-history evolution in *Drosophila melanogaster*. *Evolution* **66** 3390–3403. https://doi.org/10.1111/j.1558-5646.2012.01710.x.

SCHRAIBER, J. G., EVANS, S. N. and SLATKIN, M. (2016). Bayesian inference of natural selection from allele frequency time series. *Genetics* **203** 493–511. https://doi.org/10.1534/genetics.116.187278.

SPITZER, K., PELIZZOLA, M. and FUTSCHIK, A. (2020). Supplement to "Modifying the Chi-square and the CMH test for population genetic inference: Adapting to overdispersion." https://doi.org/10.1214/19-AOAS1301SUPPA, https://doi.org/10.1214/19-AOAS1301SUPPB, https://doi.org/10.1214/19-AOAS1301SUPPC.

STEINRÜCKEN, M., BHASKAR, A. and SONG, Y. S. (2014). A novel spectral method for inferring general diploid selection from time series genetic data. *Ann. Appl. Stat.* **8** 2203–2222. MR3292494 https://doi.org/10.1214/14-AOAS764

TAUS, T., FUTSCHIK, A. and SCHLÖTTERER, C. (2017). Quantifying selection with pool-seq time series data. *Mol. Biol. Evol.* **34** 3023–3034. https://doi.org/10.1093/molbev/msx225.

TERHORST, J., SCHLÖTTERER, C. and SONG, Y. S. (2015). Multi-locus analysis of genomic time series data from experimental evolution. *PLoS Genet.* **11** e1005069. https://doi.org/10.1371/journal.pgen.1005069.

TOBLER, R., HERMISSON, J. and SCHLÖTTERER, C. (2015). Parallel trait adaptation across opposing thermal environments in experimental *Drosophila melanogaster* populations. *Evolution* **69** 1745–1759. https://doi.org/10.1111/evo.12705.

TOBLER, R., FRANSSEN, S. U., KOFLER, R., OROZCO-TERWENGEL, P., NOLTE, V., HERMISSON, J. and SCHLÖTTERER, C. (2014). Massive habitat-specific genomic response in *D. melanogaster* populations during experimental evolution in hot and cold environments. *Mol. Biol. Evol.* **31** 364–375. https://doi.org/10.1093/molbev/mst205.

TOPA, H., JÓNÁS, Á., KOFLER, R., KOSIOL, C. and HONKELA, A. (2015). Gaussian process test for high-throughput sequencing time series: Application to experimental evolution. *Bioinformatics* **31** 1762–1770.

TURNER, T. L. and MILLER, P. M. (2012). Investigating natural variation in *Drosophila* courtship song by the evolve and resequence approach. *Genetics* **191** 633–642. https://doi.org/10.1534/genetics.112.139337.

TURNER, T. L., STEWART, A. D., FIELDS, A. T., RICE, W. R. and TARONE, A. M. (2011). Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genet.* **7** e1001336. https://doi.org/10.1371/journal.pgen.1001336.

VLACHOS, C., BURNY, C., PELIZZOLA, M., BORGES, R., FUTSCHIK, A., KOFLER, R. and SCHLÖTTERER, C. (2019). *Genome Biol.* **20** 169. https://doi.org/10.1186/s13059-019-1770-8.

WAPLES, R. S. (1989). A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121** 379–391.

WILSON, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proc. Natl. Acad. Sci. USA* **116** 1195–1200. MR3904688 https://doi.org/10.1073/pnas.1814092116