Genet. Res., Camb. (2010), **92**, *pp.* 309–320.
doi:10.1017/S0016672310000339

Testing temporal changes in allele frequencies: a simulation approach

EDSON SANDOVAL-CASTELLANOS*
Laboratorio de Genética Ecológica y Evolución, Instituto de Ecología, Universidad Nacional Autónoma de México, Circuito exterior de Ciudad Universitaria, Mexico City, P.C. 04510, Mexico

(*Received 28 December 2009 and in revised form 11 July 2010*)

Summary

Analysis of the temporal variation in allele frequencies is useful for studying microevolutionary processes. However, many statistical methods routinely used to test temporal changes in allele frequencies fail to establish a proper hypothesis or have theoretical or practical limitations. Here, a Bayesian statistical test is proposed in which the distribution of the distances among sampling frequencies is approached with computer simulations, and hypergeometric sampling is considered instead of binomial sampling. To validate the test and compare its performance with other tests, agent-based model simulations were run for a variety of scenarios, and two real molecular databases were analysed. The results showed that the simulation test (ST) maintained the significance value used ($\alpha = 0{\cdot}05$) for a vast combination of parameter values, whereas other tests were sensitive to the effect of genetic drift or binomial sampling. The differences between binomial and hypergeometric sampling were more complex than expected, and a novel effect was described. This study suggests that the ST is especially useful for studies with small populations and many alleles, as in microsatellite or sequencing molecular data.

1. Introduction

All microevolutionary processes produce changes in allele frequencies. Such processes could therefore be detected by analysing the temporal variation in allele frequencies. If this variation is too high to be caused by sampling error and genetic drift alone, it can be inferred that other evolutionary forces are responsible. The tests that are most commonly used to compare samples taken over time are the homogeneity χ^2, t, or G tests; analysis of molecular variance (AMOVA); and population divergence statistics (e.g. F_{ST}) (examples in Jorde & Ryman, 1995; Viard *et al.*, 1997; Laikre *et al.*, 1998; White *et al.*, 1998; Rank & Dahlhoff, 2002; Nayar *et al.*, 2003; Säisä *et al.*, 2003; Williams *et al.*, 2003; Han & Caprio, 2004*a*, *b*; Kollars *et al.*, 2004; Le Clerc *et al.*, 2005). These tests are not completely adequate, because their null hypothesis fails to account for genetic drift (Gibson *et al.*, 1979; Waples, 1989*a*), and this may lead to an overestimation of the significance values.

Some tests that have been developed for this purpose so far have had a limited impact, most likely because the majority were developed for specific purposes (examples in Fisher & Ford, 1947; Lewontin & Krakauer, 1973; Schaffer *et al.*, 1977; Gibson *et al.*, 1979; Wilson, 1980; Watterson, 1982; Mueller *et al.*, 1985). More recently, Goldringer & Bataillon (2004) proposed that using a simulated distribution of $F_{c,l}$ (the standardized variance in allele frequencies between generations) can be useful for testing temporal changes in allele frequencies. More recently, Bollback *et al.* (2008) developed a method for estimating N_e and s (selection coefficient) using transition probabilities whose numerical solution could be used for testing temporal changes, too. In addition, Beaumont (2003) compared several methods attaining Markov chain Monte Carlo (MCMC) algorithms and importance sampling (IS) for estimating population growth or decline, which also could be used for testing temporal changes in allele frequencies. However, those proposals did not provide the common researcher with a practical way to implement them.

* Tel: +(52)(55)56229005. e-mail: esandoval@miranda.ecologia.unam.mx

A test for general usage whose implementation was practical was developed by Waples (1989*a*), and used χ^2 adjusted to take into account the genetic drift. Nevertheless, this test may overestimate the probability values with a large number of samples and small population sizes (Waples, 1989*b*). Moreover, implementation of the test is complicated when there are many alleles and samples (Goldringer & Bataillon, 2004).

In addition, the most frequently used method to model a sample taken from a population is by binomial (replacement) sampling, but the samples should be considered strictly hypergeometric (non-replacement) (Pollak, 1983; Waples, 1989*b*) because, in reality, the sampling rarely implies the return and randomization of an organism to the population before the next organism is taken. Waples' test overcame that problem by considering the samples as drawn from the previous generation gene pool from which they represent binomial samples (Waples, 1989*a*, *b*).

Here, I present a general-use statistical test that is based on computer simulations under a Bayesian background, incorporating binomial sampling for change of generation and hypergeometric sampling for getting effective population and samples, and a user-friendly computer programme that performs the proposed analysis and the Waples test for a number of different scenarios (e.g. multiallelic systems, several samples, N and N_e variables).

2. Materials and methods

(i) *The algorithm*

Consider a locus with k alleles in a panmictic population with discrete generations of size N and effective population size N_e (number of reproductive organisms). Let $P_0, \ldots, P_t$ be the vectors of allele frequencies in the whole population at generations 0, 1, …, t, respectively, and $P_{e0}, \ldots, P_{et}$ the frequency vectors of the effective population of 0, 1, …, t generations. Let X and Y be the frequency vectors in the samples taken at generations 0 and t, with sampling sizes S_0 and S_t.

$$X=\{X_1, X_2, \ldots, X_k\};\ \sum_{i=1}^{k} X_i=1,$$

$$Y=\{Y_1, Y_2, \ldots, Y_k\};\ \sum_{i=1}^{k} Y_i=1,$$

$$P_j=\{P_{j1}, P_{j2}, \ldots, P_{jk}\};\ \sum_{i=1}^{k} P_{ji}=1,$$

$$P_{ej}=\{P_{ej1}, P_{ej2}, \ldots, P_{ejk}\};\ \sum_{i=1}^{k} P_{eji}=1,$$

where j indicates the generation and the second subindex denotes the allele.

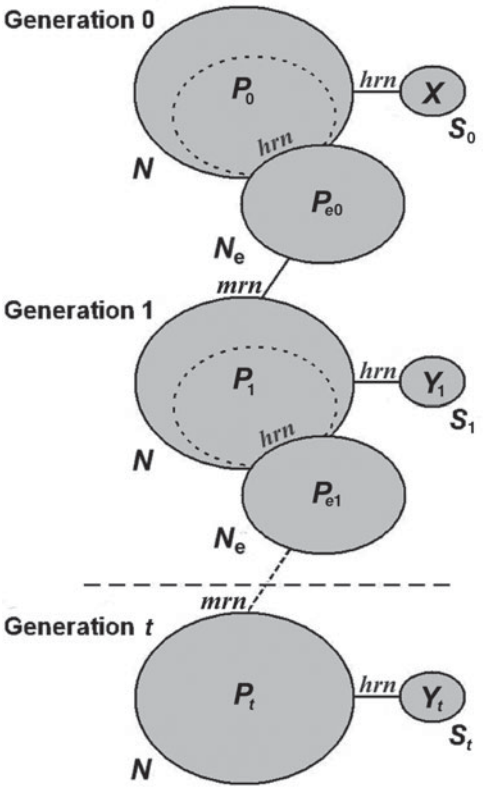

Fig. 1. Model followed by the simulations. Each ellipse represents a population. The size of the population is indicated by the term outside the ellipse, and the allele frequency of one allele at a particular locus is represented by the term inside. The subindex indicates the generation number. The effective population is a splinter group of the total population, as are the samples (they represent non-replacement samples). The total population is obtained by random mating of a very large number of gametes, and therefore, the total population can be modelled as sampling with replacement from the previous effective population. The algorithm substitutes these sampling processes by generating random numbers, whether hypergeometric (***hrn***) or multinomial (***mrn***). When the samples are separated by t generations, the intermediate non-sampled generations can be simulated only by a binomial deviate (see text).

Following Nei & Tajima's (1981) and Pollak's (1983) models, the simplest algorithm compared two samples taken from two consecutive generations as follows (see Fig. 1): after the initial parameters (N, N_e, S_0, S_t) were set, the frequencies in the samples (X and Y) and in the effective population (P_{e0}) were generated with hypergeometric multivariate random vectors, while the next generation frequencies (P_1) were obtained from multinomial deviates. After the process was repeated many times, the frequency of the occasions when the distance between the simulated allele frequencies was larger than the observed one $\left(\partial_{(\check{X},\check{Y})} \geqslant \partial_{(X_{\text{obs}}, Y_{\text{obs}})}\right)$ was taken as the P-value of the test.

The distance between m samples, X, Y_1, Y_2, …, Y_{m-1} (all taken at different times), had the

https://doi.org/10.1017/S0016672310000339 Published online by Cambridge University Press

form $\partial(X, Y_1, Y_2, \ldots Y_{m-1}) = \sum \| Y_{ji} - Y_{j-1,i} \|_1 = \sum_{j=1}^{m} \sum_{i=1}^{k-1} (Y_{ji} - Y_{j-1,i})$, where $Y_0 = X$. Notice that there is no exponent for the difference since it would make small differences even lower, distorting the magnitude of the differences.

Samples separated by t generations required an additional hypergeometric deviate (for effective population) and a multinomial deviate (for total population) for each additional generation. Alternatively, as other authors have argued, intermediate non-sampled generations could be generated only by multinomial deviates, because an effective population can be considered a multinomial draw from the previous generation (Waples, 1989*a*).

The described procedure dealt with uncertainty about N and N_e, and the allele frequencies at generation zero (P_0).

Estimation of N and N_e based on the temporally spaced samples would be redundant and clearly incorrect. Instead, two approaches can be performed, and both are suggested here: assaying of several reasonable combinations of N and N_e values by using the available knowledge about the population and employing an estimation of N_e with a method other than the temporal one, for instance, the Waples & Do (2008) and Hill (1981) linkage disequilibrium methods or Pudovkin *et al.*'s (1996) heterozygote excess method, the last two implemented in Ovenden *et al.*'s (2007) Software NeEstimator.

Uncertainty about P_0 required some mathematical treatment (see above).

(ii) *Bayesian analysis*

For representation simplicity, let us consider only the case of two temporally spaced samples X and Y. A Bayesian approach could use the predictive distribution of the allele frequencies in the samples, $\tilde{X}$ and $\tilde{Y}$, conditional to the observed frequencies:

$$f(\tilde{X}, \tilde{Y} | X_{\text{obs}}, Y_{\text{obs}}). \quad (1)$$

Over such distribution, the rejection region would be established as the one with a volume of α, where the distances between $\tilde{X}$ and $\tilde{Y}$, $\partial_{(\tilde{X}, \tilde{Y})}$ are as high as possible. A better approach could be performed by calculating the P-value by estimating the volume under the curve drawn by eqn (1), where $\partial_{(\tilde{X}, \tilde{Y})} \geqslant \partial_{(X_{\text{obs}}, Y_{\text{obs}})}$ is true. Such volume could be obtained only by integrating eqn (1), whose form should be written in terms of the unknown hyperparameter P_0:

$$f(\tilde{X}, \tilde{Y} | X_{\text{obs}}, Y_{\text{obs}}) = \int f(\tilde{X}, \tilde{Y} | P_0) f(P_0 | X_{\text{obs}}, Y_{\text{obs}}) \partial P_0. \quad (2)$$

Obtaining an analytic closed form for eqn (2) would be quite a problem, since it would require extensive and even intractable integration; however, this problem could be overcome by simulating random deviates from it. Each simulation could be done in two sequential stages: (i) simulating P_0 deviates from $f(P_0 | X_{\text{obs}}, Y_{\text{obs}})$; and (ii) simulating $\tilde{X}$ and $\tilde{Y}$ from $f(\tilde{X}, \tilde{Y} | P_0)$ for every P_0 vector obtained in the first stage, in order to finally calculate $\partial_{(\tilde{X}, \tilde{Y})}$. This algorithm would actually generate vectors $\tilde{X}$ and $\tilde{Y}$ from the desired distribution (eqn 2). The algorithm described in section (i) (also modelled in Fig. 1) indeed is valid for the second stage (ii), but the first one (i) has some potential difficulties, as the simulation from the inverse of a hypergeometric multivariate density. Two approaches were used for dealing with the simulation of P_0 from $f(P_0 | X_{\text{obs}}, Y_{\text{obs}})$: an empirical Bayes (EB) and a fully Bayesian with $f(P_0 | X_{\text{obs}}, Y_{\text{obs}})$ approached by a more friendly distribution.

EB approaches use the observed data to estimate some intermediate parameter (in this case, P_0) and then proceed as though it was a known quantity (Carlin & Louis, 2009). Thus, instead of using eqn (2), simulations of $\tilde{X}$ and $\tilde{Y}$ were drawn from

$$f(\tilde{X}, \tilde{Y} | \hat{P}_0). \quad (3)$$

where $\hat{P}_0$ is a moments' estimator of P_0 done by the method described by Waples (1989*a*). In summary, the EB algorithm consists of the following:

- Calculate $\partial_{(X_{\text{obs}}, Y_{\text{obs}})}$.
- Estimate P_0 by using the data from all the samples according to Waples (1989*a*).
- Simulate a large number of values of $\tilde{X}$ and $\tilde{Y}$ from eqn (3) by using the algorithm described in section (i) with the estimation of P_0 as the initial frequencies (the ones of the population at generation zero). For each simulation, calculate $\partial_{(\tilde{X}, \tilde{Y})}$ and record the number of times when $\partial_{(\tilde{X}, \tilde{Y})} \geqslant \partial_{(X_{\text{obs}}, Y_{\text{obs}})}$is true;
- Finally, the P-value of the test corresponds to the quotient between the number recorded in the simulations and the overall number of simulations generated.

EB procedures could give reliable estimates since the posterior still used the samples' information but at much lower computational cost (Casella, 1985).

The fully Bayesian approach still used simulations from a distribution (eqn 2), but replaces the first stage of the simulation, that is, the simulation of deviates from $f(P_0 | X_{\text{obs}}, Y_{\text{obs}})$, whose exact form would be the inverse of a hypergeometric multivariate distribution, with a more friendly one, the inverse of a multinomial distribution, namely, a Dirichlet distribution whose simulation is much simpler (Haas & Formery, 2002). The Dirichlet distribution should have the number of parameters equal to the number of alleles, so that the parameters $\alpha_1, \alpha_2, \ldots, \alpha_k$ should correspond, each one, to the average of the absolute allele frequencies

https://doi.org/10.1017/S0016672310000339 Published online by Cambridge University Press

over all the samples. However, since the frequencies at a determined sample reflect only the frequencies at generation zero, P_0, after t generations of drift, they should be adjusted to take into account such variance represented by the term $(1-1/N_e)^t$, which has already been used for analogous purposes (Waples, 1989*b*). Thus, the parameters of the Dirichlet distribution were set to $\alpha_i = S_0 X_i + S_t Y_i^*$, where $Y_i^* = Y_i(1-1/N_e)^t$.

Therefore, the algorithm for this fully Bayesian test is as follows:

- Calculate $\partial_{(X_{obs},\, Y_{obs})}$.
- Calculate the parameters of the Dirichlet distribution as indicated above.
- Simulate a large number of values of $\tilde{X}$ and $\tilde{Y}$ from an approach of eqn (2) by, first, drawing a random vector P_0 from the Dirichlet distribution described above and afterwards using the algorithm described in section (i) with the vector P_0 obtained in the previous step as the initial frequencies. Then, for each simulation, calculate $\partial_{(\tilde{X},\tilde{Y})}$ and record the number of times when $\partial_{(\tilde{X},\tilde{Y})} \geqslant \partial_{(X_{obs},\, Y_{obs})}$ is true.
- Finally, the P-value of the test corresponds to the quotient between the number recorded in the simulations and the overall number of simulations generated.

The P-value is a sensitive approach to the probability of obtaining a $\partial_{(\tilde{X},\tilde{Y})} \geqslant \partial_{(X_{obs},\, Y_{obs})}$ under pure drift, and so the P-value tests the null hypothesis of the lack of differences caused by forces in addition to gene drift.

(iii) *Relaxing some assumptions*

One advantage of the simulation test (ST) is that relaxing of some assumptions resulted straightforward, as the implementation of the two sampling schemes Nei & Tajima (1981) called Plan I and Plan II that consisted of taking organisms before or after reproduction. The difference consisted of including or not the sampled genes in the population for potentially generating the next generation.

Another advantage would be the implementation of multiple samples (all from the same population at different times) or the use of values of N and N_e changing over generations.

Such modifications also justify the use of the Nei-Tajima (1981) model since the use of the Waples (1989*b*) model (substituting hypergeometrical plus binomial sampling by binomial sampling of the previous generation) would hardly allow relaxing those assumptions without some computational drawbacks.

(iv) *Multiple loci*

For a multilocus test, it was necessary to programme the addition of the differences among simulated frequencies for all the loci and compare them with the observed frequencies. However, this approach is too demanding computationally. In addition, such an approach can be used only to determine whether the entire set of loci gives a significant result; identifying significant results for individual loci is a much more complicated task that involves multiple-hypothesis testing. Traditionally, composed hypotheses have been tested by sequential Bonferroni-type procedures as shown by Rice (1989) and others, reaching their maximum statistical power with the Benjamini & Hochberg (1995) method, which controlled the false discovery rate (FDR; the proportion of null hypothesis erroneously rejected), instead of the family-wise error rate (FWER; the probability of making one or more type I errors) as their predecessors. However, Storey (2002) showed a new approach to the problem, working on an improved FDR and q-values (analogues to P-values that reflect the proportion of false positives). They yielded many theoretical advantages and a large improvement on statistical power, being especially applicable to genetic data, as demonstrated in the work of Storey & Tibshirani (2001), who presented a practical way to implement the analysis, which is highly recommended. In addition, this topic has received a huge amount of interest in recent years as the methods to estimate the FDR have multiplied (e.g. Strimmer, 2008) as well as the applications to genome-wise studies (e.g. Forner *et al*., 2008).

(v) F_T*: a statistic for quantifying differences among temporal samples*

Some researchers have quantified the temporal differentiation by estimating F_{ST} from samples taken from the same population at different times as if the samples were from different geographically separated locations. The method is clearly erroneous, but the necessity of quantifying (and not only testing) the differentiation accumulated in time is real. Wright's F_{ST} indicates subpopulation differentiation by an uncoupling of heterozygosis between the overall population and subpopulations levels. This departure of heterozygosis was thought to increase with the time by gene drift when gene flow is restricted among subpopulations. With some caution, F_{ST} could be useful for the quantification of temporal differences, taking into account that F_{ST} was designed to account for the differentiation generated only by gene drift. Nevertheless, researchers in general are actually not interested in quantifying gene drift but other evolutionary forces, so what is here proposed is the use of a statistic analogue to F_{ST} easily interpretable as reflecting differentiation equivalent to the obtained value of an F_{ST}. Such statistic is applicable to samples taken from the same population (the same geographic location) at different times and consists of an F_{ST} from which

an average-gene-drift term has been subtracted. This average-gene-drift term is the average of F_{ST} estimations among temporally simulated samples and is here named $\bar{F}^s$. Since that average gene drift behaves as a random variable, it gives the advantage of estimating at one time the value and statistical significance by recording the frequency of occurrences when the F_{ST} among simulated samples was smaller than the F_{ST} estimated from real samples. The resulting statistic here called F_T accounts the differences among samples taken at different times after subtracting the mean gene drift, i.e. due to other evolutionary forces. In the programme, F_T was calculated as $F_T = F_{ST} - \bar{F}^s$, while F_{ST} was calculated as Nei's (1973) formula: $F_{ST} = (H_T - H_S)/H_T$, where H_S is the average Hardy–Weinberg heterozygosity and $H_T = 1 - \sum \bar{p}_i^2$ for any number of alleles.

(vi) *Validation of the test*

Agent-based model (ABM) simulations were programmed for playing the role of real populations where virtual samples were taken from, and used to evaluate the effectiveness and discover the properties of four tests: the above-described ST, the proposed test of significance of F_T, the adjusted Waples test (WT), and a conventional χ^2 contingency test (ChT), with different scenarios and combinations of parameters. Unlike ST simulations, the ABM simulations represented detailed populations where organisms were represented and sampled one by one, the populations mated randomly in a process where alleles were passed randomly to the offspring, and the number of descendants was obtained with a Poisson distribution. For each combination of parameters, 10 000 ABM simulations were run, and the number of significant tests (with $\alpha = 0{\cdot}05$) were recorded. In addition, several assays were run with a modified ST whose samples were all binomial (multinomial for several alleles) instead of hypergeometric, in order to assess the potential processes affecting allele frequencies whose effect would be neglected by binomial sampling. In addition, statistical robustness was assessed by performing runs in which incorrect population sizes were used (i.e. the population sizes of the ABM simulations and those used in the tests were different), while the statistical power was examined by introducing a constant increase in the frequency of one allele emulating a positive natural selection process.

Finally, the four tests were applied to two real datasets: (i) frequencies of four microsatellite loci from the snail *Bulinus truncatus* (Viard *et al.*, 1997) taken at three different times separated by four and one generations, which were obtained from 12 locations, and (ii) frequencies from six sequenced genes (analysed as biallelic systems) that are involved in the expression of the skin colour of horses, dated at 5 times over a 12 900-year period (13 100, 3700, 2800, 600, and 200 BC) reported by Ludwig *et al.* (2009) (Table 1).

(vii) *Programming*

Multinomial and hypergeometric multivariate deviates were generated by iterative simulation from their univariate marginal densities (Haas & Formery, 2002; Gelman *et al.*, 2004), which were obtained with Kemp's (1986) and Voratas & Schmeiser's (1985) methods, respectively. Dirichlet deviates were obtained as explained in Gelman *et al.* (2004) and in Haas and Formery (2002) by using the random gamma generator contained in the module 'random' (A. Miller, available at: http://www.Mathtools.net). For uniform deviates, the RANLUX module was used (Lüscher, 1994; James, 1994).

3. Results

(i) *Effect of the parameters*

After more than 400 ABM simulations were run (lasting more than 2000 h of computer time), the results with the parameters that caused the more pronounced effects are shown below.

For most combinations of the parameters, the ST, F_T test, and WT showed small deviations from the expected 5% of the significant tests. However, the ChT gave higher numbers on the significant tests, which apparently increased as the effects of genetic drift accumulated, for the lower population sizes (Fig. 2*A*), the higher numbers of generations (Fig. 2*B*), or higher numbers of samples (Fig. 2*C* and *D*). Moderate overestimation of the proportion of the significant tests was observed for the WT with small population sizes (<500) (Fig. 2*A*), high numbers of generations (Fig. 2*B*) or high numbers of samples (Fig. 2*C* and *D*). The proportion of significant tests obtained with the WT was greater than 6% only with very low population sizes (Fig. 2*A*) or with numerous samples (>10) (Fig. 2*D*). The low-population-sizes departure could be related to a high sample size to effective population size ratios (S/N_e) in addition to the small population size effect. The effect of S/N_e has been demonstrated by Waples (1989*a*) and was also detected here in a group of simulations with different S/N_e ratios (from 0·1 to 0·9; results not shown). Interestingly, the ChT and WT performed well with high numbers of alleles or loci (not shown), but both showed a higher proportion of significant tests than the ST or F_T test.

On the other hand, for the ST and F_T test, differences greater than 2% from the expected value of 5% of the significant tests were not observed for most combinations of the parameters, and in general, the

https://doi.org/10.1017/S0016672310000339 Published online by Cambridge University Press

Table 1. *Results for the two real data sets analysed. In the snail data set, the numbers under the label 'No. of generations between samples' indicate the number of generations between the first and last samples, while the numbers inside the parentheses indicate the generations between consecutive samples. Grey cells indicate significant results* ($\alpha = 0{\cdot}05$)*, and * indicates analyses that were not possible because of a lack of data or because the locus was monomorphic*

Microsatellites of the snail (*Bulinus truncatus*)

Location	No. of samples	Locus	No. of alleles	No. of generations between samples	Probability values of the tests			
					ST	Ft	WT	ChT
Boyze I			2	5(4/1)	0·5407	0·5230	0·1337	0·1670
Boyze II			1	5(4/1)	*	*	*	*
Doubalma			1	5(4/1)	*	*	*	*
Bala			2	5(4/1)	0·4254	0·4000	0·2369	0·1063
Kobouri			1	5(4/1)	*	*	*	*
Tera R	3	BT1	2	5(4/1)	0·9776	0·9670	0·9782	0·9774
Tera D			3	5(4/1)	0·0809	0·0860	0·0048	0·0002
Namaga PM			3	5(4/1)	0·1828	0·1921	0·1437	0·1109
Namaga B			4	5(4/1)	0·2957	0·3120	0·0897	0·0551
Namaga W			3	5(4/1)	0·0509	0·0521	0·0000	0·0000
Mari Sud			3	5(4/1)	0·4015	0·5120	0·2108	0·1931
Mari Nord			3	5(4/1)	0·4234	0·4530	0·5404	0·5681
Boyze I			1	5(4/1)	*	*	*	*
Boyze II			1	5(4/1)	*	*	*	*
Doubalma			2	5(4/1)	0·0037	0·0100	0·0079	0·0088
Bala			1	5(4/1)	*	*	*	*
Kobouri			1	5(4/1)	*	*	*	*
Tera R	3	BT6	5	5(4/1)	0·0014	0·0100	0·0001	0·0001
Tera D			5	5(4/1)	0·0034	0·0120	0·0000	0·0000
Namaga PM			3	5(4/1)	0·1006	0·1110	0·0473	0·0170
Namaga B			3	5(4/1)	0·0913	0·0920	0·0776	0·0358
Namaga W			2	5(4/1)	0·0719	0·0750	0·0054	0·0011
Mari Sud			4	5(4/1)	0·1553	0·1550	0·0054	0·0038
Mari Nord			5	5(4/1)	0·1137	0·1000	0·0509	0·0242
Boyze I			5	5(4/1)	0·0004	0·0010	0·0000	0·0000
Boyze II			4	5(4/1)	0·0939	0·1001	0·2467	0·2139
Doubalma			2	5(4/1)	0·0331	0·0340	0·0330	0·0350
Bala			5	5(4/1)	0·5207	0·4900	0·0140	0·0248
Kobouri			4	5(4/1)	0·0115	0·0100	0·0022	0·0015
Tera R	3	BT12	14	5(4/1)	0·0001	0·0000	0·0606	0·1027
Tera D			12	5(4/1)	0·0301	0·0310	0·0000	0·0001
Namaga PM			12	5(4/1)	0·1270	0·1300	0·0073	0·0009
Namaga B			12	5(4/1)	0·0000	0·0000	0·0053	0·0382

Genes for horse coat colour

Comparison	No. of samples	Locus	No. of alleles	No. of generations between samples	Probability values of the tests			
					ST	Ft	WT	ChT
−13100 versus −3700				9400	0·0003	0·0002	0·0001	0·0000
−3700 versus −2800	2	ASIP	2	900	0·4080	0·3528	0·3286	0·3112
−2800 versus 600				2200	0·2591	0·2635	0·2437	0·1883
−600 versus −200				400	0·6356	0·6356	0·6337	0·6219
All	5			12900	0·0087	0·0017	0·0007	0·0001
−13100 versus −3700				9400	0·4598	0·4113	0·2912	0·2885
−3700 versus −2800	2	MC1R	2	900	0·0669	0·0522	0·0410	0·0375
−2800 versus −600				2200	0·6923	0·6921	0·6797	0·6435
−600 versus −200				400	0·0276	0·0295	0·0247	0·0200
All	5			12900	0·1882	0·0240	0·0000	0·0000
−13100 versus −3700				9400	*	*	*	*
−3700 versus −2800	2	KIT13	2	900	*	*	*	*
−2800 versus −600				2200	0·3630	0·2770	0·2850	0·2831
−600 versus −200				400	0·4728	0·5045	0·4451	0·4336
All	5			12900	0·5508	0·6478	0·0007	0·1750
−13100 versus −3700				9400	*	*	*	*
−3700 versus −2800	2	KIT16	2	900	0·8701	0·7840	0·3117	0·3112
−2800 versus −600				2200	0·9998	0·9984	0·9371	0·9295
−600 versus −200				400	0·1899	0·1330	0·1687	0·1408
All	5			12900	0·4114	0·3562	0·0127	0·3545
−13100 versus −3700				9400	*	*	*	*
−3700 versus −2800	2	MATP	2	900	*	*	*	*
−2800 versus −600				2200	*	*	*	*
−600 versus −200				400	0·1196	0·0840	0·0857	0·0852
All	5			12900	0·5678	0·5165	0·0000	0·0949
−13100 versus −3700				9400	*	*	*	*
−3700 versus −2800	2	SILV9	2	900	*	*	*	*
−2800 versus −600				2200	0·8692	0·8235	0·4529	0·4520
−600 versus −200				400	0·5472	0·4908	0·3340	0·3009
All	5			12900	0·4814	0·4698	0·0172	0·5921

Namaga W	10	5(4/1)	0·1711	0·1775	0·0026	0·0001	
Mari Sud	18	5(4/1)	0·1424	0·1350	0·0037	0·0016	
Mari Nord	16	5(4/1)	0·0306	0·0310	0·0002	0·0002	
Boyze I	3	5(4/1)	0·0932	0·1000	0·0026	0·0005	
Boyze II	4	5(4/1)	0·5385	0·5450	0·7900	0·7603	
Doubalma	5	5(4/1)	0·6320	0·6330	0·5956	0·5995	
Bala	4	5(4/1)	0·0842	0·0850	0·2887	0·2656	
Kobouri	5	5(4/1)	0·0639	0·0640	0·0140	0·0106	
Tera R	15	5(4/1)	0·0000	0·5000	0·4894	0·4924	
Tera D	11	5(4/1)	0·0000	0·4810	0·4706	0·4946	
Namaga PM	24	5(4/1)	0·0000	0·0001	0·0010	0·0053	
Namaga B	26	5(4/1)	0·0002	0·0001	0·0002	0·0016	
Namaga W	16	5(4/1)	0·0083	0·0090	0·0061	0·0739	
Mari Sud	15	5(4/1)	0·0016	0·0500	0·0059	0·0082	
Mari Nord	21	5(4/1)	0·0000	0·3610	0·0351	0·0605	

3

BT13

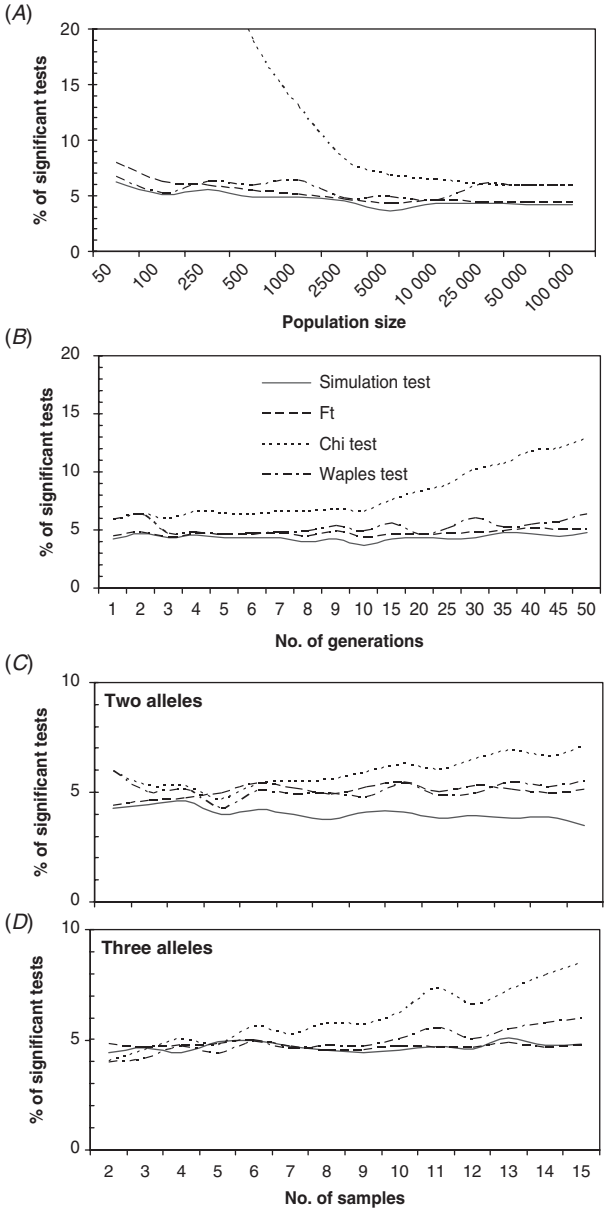

Fig. 2. Percentage of significant tests obtained for different values of: (*A*) total population sizes (N), (*B*) numbers of generations between samples; and an increasing number of samples/generations, with two alleles (*C*) or three alleles (*D*). The default settings were (unless otherwise is indicated) as follows: $N = 10\,000$, $N_e = N/2$, $t = 5$, two samples with sizes $S_0 = S_5 = 100$, number of alleles $k = 2$, initial allele frequencies of 0·5, sampling Plan II and fully Bayesian algorithm. In (*C*) and (*D*), only one generation separated consecutive samples, and the initial allele frequencies were 0·95/0·05 and 0·7/0·28/0·02, respectively.

proportion of significant tests obtained with these methods was below 5%.

In order to determine whether the small departures from the expected number of significant tests that were found for several parameter ranges were accumulative, a set of ABM simulations was run with 'problematic' combinations of parameter values.

https://doi.org/10.1017/S0016672310000339 Published online by Cambridge University Press

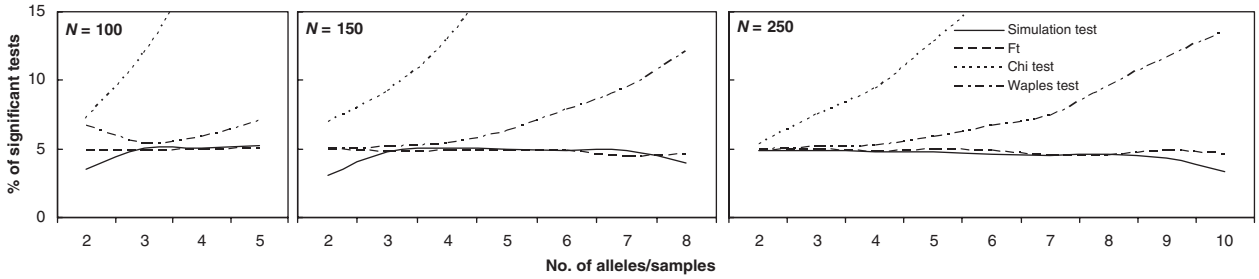

Fig. 3. The graphics show the percentages of significant values obtained for different numbers of samples and alleles and for three different population sizes, N, as indicated in the graphics. The higher numbers of alleles/samples could not be used with the smaller population sizes ($N=100$ and 150), because the likelihood of an allele being lost is very high with a large number of alleles and a small number of organisms (the general simulations do not allow samples with missing alleles). The number of samples was the same as the number of alleles for each run. They were increased simultaneously just to assess the combined effect of the parameters in a simple form, because our goal was not to analyse the effect of each parameter (that was already done) but the magnitude the departures can reach with certain 'problematic' combinations of parameters. For that purpose, the following settings were used: $N_e=N/2$, one generation between consecutive samples, sample sizes were all the same $S_i=20$, sampling Plan II and fully Bayesian algorithm. The initial allele frequencies were as follows: 0·5/0·5 (two alleles), 0·25/0·25/0·5 (three alleles), 0·167/0·167/0·167/0·5 (four alleles), 0·125/0·125/0·125/0·125/0·5 (five alleles), 0·1/0·1/0·1/0·1/0·1/0·5 (six alleles), 0·083/…/0·083/0·5 (seven alleles), 0·071/…/0·071/0·5 (eight alleles), 0·0625/…/0·0625/0·5 (nine alleles) and 0·055/…/0·055/0·5 (ten alleles).

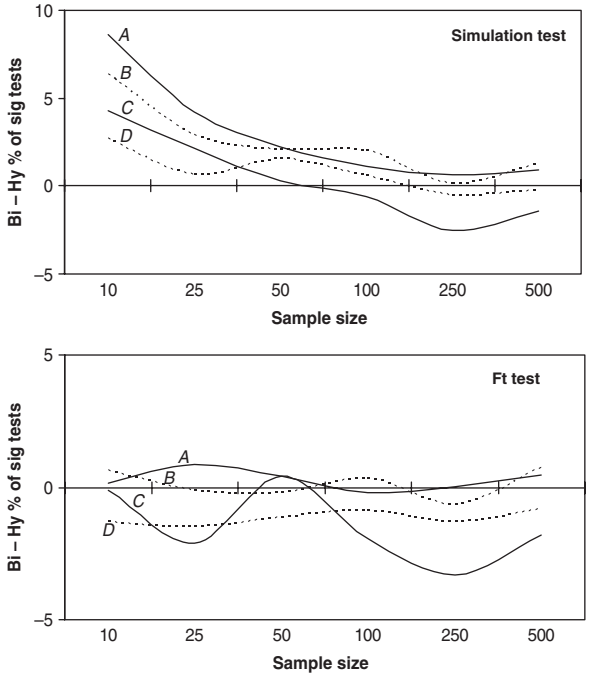

Fig. 4. Difference between binomial and hypergeometric significant STs, and for tests over the F_T statistic. Graphics show the % of the significant tests obtained with binomial ST minus the % obtained with hypergeometric ST, as a function of sample sizes. Two S–N_e ratios and sampling plans were used, $S/N_e=0·9$ and $S/N_e=0·5$; and Plan I (before reproduction) and Plan II (after reproduction). Curves correspond to: (*A*) $S/N_e=0·9$, Plan II; (*B*) $S/N_e=0·5$, Plan II; (*C*) $S/N_e=0·9$, Plan I, (*D*) $S/N_e=0·5$, Plan I. A curve in the positive region meant that the % of significant bSTs was larger than the % of hypergeometric tests, and a curve in the negative region the converse. The default settings were as follows: $N=2N_e$, $t=5$ (generations between samples), $k=2$, initial allele frequencies of 0·5/0·5 and fully Bayesian algorithm. Notice that, since N, N_e were fixed for each curve, the x-axis not only indicates increasing sample sizes but also increasing N and N_e.

Figure 3 shows that when many alleles were analysed, with some at very low frequencies, and the samples were increased simultaneously while the population size was reduced, the proportion of significant tests obtained with the ChT rose rapidly to very high values. In addition, the otherwise stable WT yielded a proportion of significant tests that was greater than 10%. The ST and F_T test maintained their previously observed stability, giving a proportion of significant tests which was close to 5%.

(ii) *Binomial simulation test* (*bST*)

After many combinations of parameters were assayed, the bST presented relevant increases of significant tests with a decreasing number of alleles, a decreasing number of samples, low sample and population sizes, an increasing number of generations and values of S/N_e closer to one (not shown). The number of significant tests obtained with the bST reached 15%. In order to inquire into this unexpected result, a set of runs with small population sizes, decreasing sample sizes, different S/N_e values as well as different sampling plans were run. Figure 4 displays the differences of the percentage of statistical tests between bST and (hypergeometric) ST, which showed positive values when the binomial version had a higher % and negative when the hypergeometric version had a higher %. The curves corresponding to two different S/N_e values and the two sampling plans support that two putative causes explain the binomial test bias: (i) a reduction in the total population and (ii) a negative correlation between population and sample allele frequencies. Both produced the effect of enhancing the differences among the frequencies of the samples when hypergeometric and sampling

https://doi.org/10.1017/S0016672310000339 Published online by Cambridge University Press

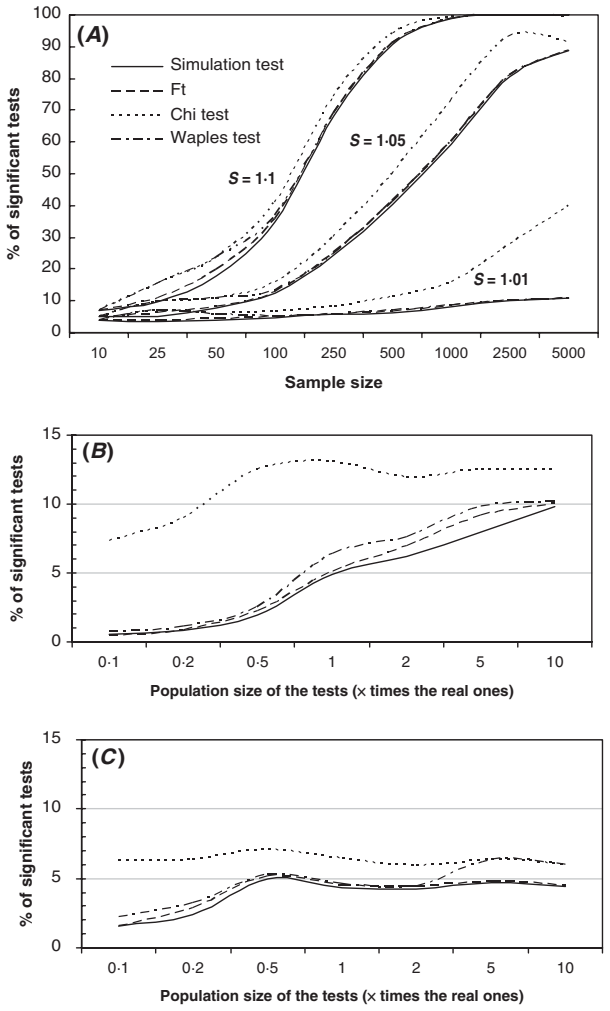

Fig. 5. (*A*) Percentage of significant tests obtained when the frequency of one allele in the 'real population' (the ABM simulation) was increased by a given amount ($s = 1{\cdot}01\times$, $1{\cdot}05\times$ and $1{\cdot}1\times$) each generation, emulating a positive selection process. (*B*, *C*) Percentage of significant tests when incorrect values of N were used by the tests (x-axis) for two different values of real population size (the used in general simulations): $N = 1000$ (*B*) and $N = 10\,000$ (*C*). The default parameters used were as follows: $N = 10\,000$, $N_e = N/2$, $t = 5$, $S_0 = S_5 = 100$, number of alleles $k = 2$, initial allele frequencies of 0·5, sampling Plan II and full Bayesian algorithm.

Plan II was present, making the binomial test to overestimate significant tests by simulating samples with lower differences than the real ones. This effect was present only with a combination of low population sizes, high S/N_e values and sampling Plan II (see the complete explanation in section 4).

(iii) *Statistical robustness and power*

In the analysis of statistical robustness, as the population size used by the tests was increased with respect to the real population size (ABM simulations), the number of significant tests increased, while a reduction in the sizes used by the tests produced a reduction in the number of significant tests. As expected, the ChT appeared more robust, because it did not require the parameters N and N_e (Fig. 5*B* and *C*). The statistical power of the ChT was greater than that of the other tests for a selection coefficient of 1·01, but similar to that of the other tests for higher selection coefficients (Fig. 5*A*). The ST, F_T test and WT all showed a very similar statistical power to each other with each of the three different selection coefficients used.

(iv) *Real data sets*

Table 1 shows the results for the two real data sets. In the data set of the snail *B. truncatus*, from 41 tests on microsatellite loci that contained between 2 and 26 alleles, around 12 presented quite different probability values among the tests. The loci with large differences (among the tests) were mostly the ones with many alleles at very low frequencies. ST and F_T tended to present higher values of probability than WT and ChT.

In the genes data set for equine colour coat, ST and F_T were found to tend to yield higher probabilities than WT and ChT. From the 17 tests applied, two showed significant results for all tests, and one was significant for WT and ChT ($P \sim 0{\cdot}04$) but not with the ST and F_T ($P \sim 0{\cdot}06$). The overall tests with all the samples involved simultaneously showed low probabilities with WT or ChT in five (from six) loci, while the same loci showed high probabilities (>0·2) with ST and F_T. The most relevant differences were found among tests that attained very low-frequency alleles.

4. Discussion

The observed deviations of the ChT from the expected value of 5% of the significant tests (with almost all the parameters assayed) can be explained mainly by the effect of genetic drift, as pointed out by Waples (1989*a*). Here, low population sizes, high numbers of samples and generations between samples produced an extremely high number of false significant tests with the ChT.

Contrastingly, the ST, F_T test and WT performed homogeneously, showing low to moderate deviations from the expected value. The differences were larger for the smaller population sizes, higher number of samples and higher number of loci. These deviations had different origins.

The deviations that were obtained with the WT for small population sizes may have originated in the departure of the statistic used from a χ^2 distribution, already identified by Goldringer & Bataillon (2004) who suggested the simulation-generated distribution of the $F_{c,l}$ statistic as the basis of a test for assaying temporal changes. The effect of small population sizes, together with large sample sizes, has already

https://doi.org/10.1017/S0016672310000339 Published online by Cambridge University Press

been recognized as an analytical problem that produces errors in the estimation of effective population size by the temporal method (Nei & Tajima, 1981; Pollak, 1983; Waples, 1989*b*). However, a component of this bias could be the effect of the S/N_e ratio demonstrated by Waples (1989*a*) and others that increase as population sizes decrease for constant sample sizes, and that could be related to the increased statistical power of samples proportionally larger with respect to the effective population size. In spite of non-replacement (hypergeometric) samples having a smaller variance than replacement (binomial) samples, it is expected that an underestimation of significant tests with the bST but not with the WT since the sampling used by Waples (1989*a*, *b*) refers to the previous generation and those differences are not expected to affect the results of the test, as actually happened. However, in runs with small population sizes and sample sizes similar to N_e ($S/N_e \sim 1{\cdot}0$), the bST not only underestimated the % of the significant tests, but also presented important overestimations of those percentages. This was caused by the use of sampling Plan II over most of the simulations of the study since this plan's results were more realistic. With this scenario (small N and $S/N_e \sim 1{\cdot}0$), two effects would enhance differences among allele frequencies at temporally spaced samples in a real population (with hypergeometric-like sampling): (i) since the sampling would be proportionally large and the population size small, the sampled alleles would leave the population with a size even smaller and consequently more susceptible to changes by drift and (ii) with such a small population size and proportionally large sample size, the allele frequencies in the sample and in the population (after sampling) would establish a trade-off. For instance, in a population with an allele at a frequency of 0·5 before sampling, if N were so small that $S = N/2$ is true, then the frequency of the allele of 0·7 in the sample would mean a frequency in the population of 0·3; that is, the higher the number of alleles in the first sample, the lower the remaining in the population and thus at posterior samples. Both effects are absent from binomial simulations since (i) a constraint in binomial distribution prevents programming the discounting of sampled alleles from the total population (e.g. the binomial sample could actually have more alleles of one type than the original number of alleles in the population, and so the frequency in the population after discounting the sample would be negative!) and (ii) since the alleles sampled binomially are sampled with replacement, the frequencies cannot establish a trade-off with the frequencies in the population.

Those mechanisms explain that under small population sizes, $S/N_e \sim 1{\cdot}0$, and sampling Plan II, the hypergeometric simulations as well as the real populations should present higher differences in allele frequencies among the samples than binomial simulations. Thus, for fixed (observed) frequencies, the binomial test generates fewer simulations with differences equal or higher than the observed ones, underestimating the P-values and thus overestimating the % of the significant tests. The results shown in Fig. 4 support the explained mechanisms with three facts: (i) the curves had lower values for higher sample sizes, which meant not higher S/N_e but higher population sizes, i.e. for higher population sizes, the binomial bias was lost; (ii) the curves with $S/N_e = 0{\cdot}5$ had lower values than with $S/N_e = 1{\cdot}0$, i.e. the binomial bias became weaker as the sample sizes became lower than the population sizes and (iii) the curves obtained with sampling Plan I returned to negative values when the sample sizes (i.e. the population sizes) increased, which means that in the absence of sampling Plan II the binomial bias could be inverted, and the differences would be caused by the mentioned difference between hypergeometric and binomial variances.

Thus, the Waples test, the Goldringer and Bataillon test and other binomial sampling-based tests potentially would have two sources of bias with opposite effects: the difference between the hypergeometric and binomial variances and the one explained above.

Furthermore, the WT and binomial bias observed for the increased number of generations and samples could be related to the assumption that the model of F (the standardized temporal variance in allele frequencies) has to be proportional to $t/2N_e$ which becomes particularly relevant at large t and very small N_e values (Waples, 1989*a*, *b*).

The results of the analysis of the real data sets agreed with the ABM simulation results. In the case of the snail, *B. truncatus*, the largest differences among ST/F_T and WT/ChT were found in loci with many alleles from which several had low frequencies. In addition, the study attained low population sizes ($N = 1000$) and three samples. All these factors brought the studied system closer to that represented by the 'problematic' combinations of parameters that were assayed with the ABM simulations, where the WT and ChT gave smaller P-values, in agreement with the tendency of WT and ChT to yield lower P-values than ST and F_T in the snail data set.

In addition, the second real data set analysed the frequencies of genes for horse coat colour had features similar to the problematic combinations, such as having a very large number of generations among samples, many samples and low-frequency alleles. They also showed a similar tendency: lower P-values with WT and ChT than with ST and F_T.

The possible explanation for the behaviour observed with problematic combinations is probably composed. Some components could be the mentioned departures of the χ^2 distribution or the binomial-hypergeometric bias(es). Another relevant component

https://doi.org/10.1017/S0016672310000339 Published online by Cambridge University Press

is a boundary effect of low initial allele frequencies. Additional simulations were performed to probe the effect of low allele frequencies and showed a tendency to yield increased probability values, which were larger with the ST. The boundary conditions already detected by Waples (1989*a*) involved a large probability that the allele detected was lost. Such a loss would prevent further change in the allele frequency, reducing the possibility of a type II error (Liukart *et al.*, 1999; Goldringer & Bataillon, 2004).

Another interesting and unexpected result was found with the snail data set. Five loci presented substantially larger probability values with the ChT and WT than with the ST and F_T test. These loci all corresponded to a high number of alleles, many of which had very low frequencies, and some alleles had been lost in one or more samples. One hypothesis to explain this phenomenon is that the absence of an allele in some samples, in conjunction with its occurrence at a high frequency in another sample, will occur very infrequently by chance alone, due to the mentioned boundary effect. Thus, the STs (ST and F_T), as real populations, were reluctant to allow alleles at very low frequencies arising suddenly by chance, whereas ChT and WT summarize the overall differences among alleles and samples without discriminating among common patterns or very improbable patterns (e.g. absence/presence-at-high-frequency/absence of an allele). Additional (ABM) simulations with the selection-like process acting on low-frequency alleles supported this statement.

In addition, the ST, F_T test and WT showed similar statistical robustness and statistical power, which precluded the possibility that the differences observed with the snail data were due to the effect of differential statistical powers for the tests, and not to population sizes.

Finally, it can be said that the WT, ST and F_T test performed better than the ChT, as expected, because they took into account the effect of genetic drift. However, for certain combinations of parameters, the ST and F_T test yielded more reliable results, which makes them more suitable for studies that involve many samples, low population sizes and high numbers of alleles (as in the case of microsatellite or sequence data). Further studies are required to explore the combined effect of small population sizes, multiple alleles and samples, S/N_e, together with the effect of having low-frequency alleles.

Software implementing the described tests is available at: http://sites.google.com/site/egenevol/home or from esandoval@miranda.ecologia.unam.mx by request.

I thank Manuel Uribe-Alcocer, Lourdes Barbosa-Saldaña, Luis Medrano-González, Robin Waples, and two anonymous reviewers for their comments on earlier versions of this paper.

References

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* (*Statistical Methodology*) **57**, 289–300.

Bollback, J., York, T. L. & Nielsen, R. (2008). Estimation of $2N_e$s from temporal allele frequency data. *Genetics* **179**, 497–502.

Carlin, J. B. & Louis, T. A. (2009). *Bayesian Methods for Data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *American Statistician* **39**, 83–87.

Fisher, R. A. & Ford, E. B. (1947). The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. *Heredity* **1**, 143–174.

Forner, K., Lamarine, M., Guedj, M., Dauvillier, J. & Wojcik, J. (2008). Universal false discovery rate estimation methodology for genome-wide association studies. *Human Heredity* **65**, 183–194.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Gibson, J. B., Lewis, N., Adena, M. A. & Wilson, S. R. (1979). Selection for ethanol tolerance in two populations of *Drosophila melanogaster* segregating alcohol dehydrogenase allozymes. *Australian Journal of Biological Sciences* **32**, 387–398.

Goldringer, I. & Bataillon, T. (2004). On the distribution of temporal variations in allele frequency: consequences for the estimation of effective population size and the detection of loci undergoing selection. *Genetics* **168**, 563–568.

Haas, A. & Formery, P. (2002). Uncertainties in facies proportion estimation I. theoretical framework: the Dirichlet distribution. *Mathematical Geology* **34**, 679–702.

Han, Q. & Caprio, M. A. (2004*a*). Evidence from genetic markers suggests seasonal variation in dispersal in *Heliothis virescens* (Lepidoptera: Noctuidae). *Environmental Entomology* **33**, 1223–1231.

Han, Q. & Caprio, M. A. (2004*b*). Temporal and spatial patterns of allelic frequencies in cotton bollworm (Lepidoptera: Noctuidae). *Environmental Entomology* **31**, 462–468.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetics Research* **38**, 209–216.

James, F. (1994). RANLUX: a Fortran implementation of the high-quality pseudorandom number generator. *Computer Physics Communications* **79**, 111–114.

Jorde, P. E. & Ryman, N. (1995). Temporal allele frequency change and estimation of effective size in populations with overlapping generations. *Genetics* **139**, 1077–1090.

Kemp, C. D. (1986). A modal method for generating binomial variables. *Communications in Statistics – Theory and Methods* **15**, 805–813.

Kollars, P. G., Beck, M. L., Mech, S. G., Kennedy, P. K. & Kennedy, M. L. (2004). Temporal and spatial genetic variability in the white-tailed deer (*Odocoileus virginianus*). *Genetica* **121**, 269–276.

Laikre, L., Jorde, P. E. & Ryman, N. (1998). Temporal change of mitochondrial DNA haplotype frequencies and female effective size in a brown trout (*Salmo trutta*) population. *Evolution* **52**, 910–915.

Le Clerc, V., Bazante, F., Baril, C., Guiard, J. & Zhang, D. (2005). Assessing temporal changes in genetic diversity of maize varieties using microsatellite markers. *Theoretical and Applied Genetics* **110**, 294–302.

Lewontin, R. C. & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195.

Liukart, G., Cornuet, J. M. & Allendorf, J. M. (1999). Temporal changes in allele frequencies provide estimates of population bottleneck size. *Conservation Biology* **13**, 523–530.

Ludwig, A., Pruvost, M., Reissmann, M., Benecke, N., Brockmann, G. A., Castaños, P., Cieslak, M., Lippold, S., Llorente, L., Malaspinas, A.-S., Slatkin, M. & Hofreiter, M. (2009). Coat color variation at the beginning of horse domestication. *Science* **324**, 485.

Lüscher, M. (1994). A portable high-quality random number generator for lattice field theory simulations. *Computer Physics Communications* **79**, 100–110.

Mueller, D. L., Wilcox, A. B., Ehrlich, R. P., Heckel, G. D. & Murphy, D. D. (1985). A direct assessment of the role of genetic drift in determining allele frequency variation in populations of *Euphydryas editha*. *Genetics* **110**, 495–511.

Nayar, J. K., Knight, J. W. & Munstermann, L. E. (2003). Temporal and geographic variation in *Culex pipiens quinquefasciatus* (Diptera: Culicidae) from Florida. *Journal of Medical Entomology* **40**, 882–889.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences, USA* **70**, 3321–3323.

Nei, M. & Tajima, F. (1981). Genetic drift and estimation of effective population size. *Genetics* **98**, 625–640.

Ovenden, J., Peel, D., Street, R., Courtney, A. & Hoyle, S., *et al*. (2007). The genetic effective and adult census size of an Australian population of tiger prawns (*Penaeus esculentus*). *Molecular Ecology* **16**, 127–138.

Pollak, E. (1983). A new method for estimating the effective population size from allele frequency changes. *Genetics* **104**, 531–548.

Pudovkin, A. I., Zaykin, D. V. & Hedgecock, D. (1996). On the potential for estimating the effective number of breeders from heterozygote excess in progeny. *Genetics* **144**, 383–387.

Rank, N. E. & Dahlhoff, E. P. (2002). Allele frequency shifts in response to climate change and physiological consequences of allozyme variation in a montane insect. *Evolution* **56**, 2278–2289.

Rice, W. (1989). Analyzing tables of statistical tests. *Evolution* **43**, 223–225.

Säisä, M., Koljonen, M. & Tähtinen, J. (2003). Genetic changes in Atlantic salmon stocks since historical times and the effective population size of along-term captive breeding programme. *Conservation Genetics* **4**, 613–617.

Schaffer, H. E., Yardley, D. & Anderson, W. W. (1977). Drift or selection: a statistical test of gene frequency change over generations. *Genetics* **87**, 371–379.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* (*Statistical Methodology*) **64**, 479–498.

Storey, J. D. & Tibshirani, R. (2001). Statistical significance for genomewide studies. *Proceeding of the National Academy of Sciences, USA* **100**, 440–9445.

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics* **9**, 303.

Viard, F., Justy, F. & Jarne, P. (1997). Population dynamics inferred from temporal variation at microsatellite loci in the selfing snail *Bulinus truncatus*. *Genetics* **146**, 973–982.

Voratas, K. & Schmeiser, B. (1985). Computer generation of hypergeometric random variates. *Journal of Statistical Computation and Simulation* **22**, 127–145.

Waples, R. S. (1989a). Temporal variation in allele frequencies: testing the right hypothesis. *Evolution* **43**, 1236–1251.

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

Waples, R. S. & Do, C. (2008). LDNE: a program for estimating effective populate size from data on linkage disequilibrium. *Molecular Ecology Resources* **8**, 753–756.

Watterson, G. A. (1982). Testing selection at a single locus. *Biometrics* **38**, 323–331.

White, S. E., Kennedy, P. K. & Kennedy, M. L. (1998). Temporal genetic variation in the raccoon *Procyon lotor*. *Journal of Mammalogy* **79**, 747–754.

Williams, C. L., Blejwas, K., Johnston, J. J. & Jaeger, M. M. (2003). Temporal genetic variation in a coyote (*Canis latrans*) population experiencing high turnover. *Journal of Mammalogy* **84**, 177–184.

Wilson, S. R. (1980). Analyzing gene-frequency data when effective population size is finite. *Genetics* **95**, 489–502.