THE ROYAL SOCIETY

Received 26 October 2003
Accepted 1 December 2003
Published online 18 March 2004

Allele excess at neutrally evolving microsatellites and the implications for tests of neutrality

Christian Schlötterer*, **Max Kauer and Daniel Dieringer**

Institut für Tierzucht und Genetik, Veterinärmedizinische Universität Wien, Josef Baumann Gasse 1, 1210 Wien, Austria

Skews in the observed allele-frequency spectrum are frequently viewed as an indication of non-neutral evolution. Recent surveys of microsatellite variability have used an excess of alleles as a statistical approach to infer positive selection. Using neutral coalescent simulations we demonstrate that the mean numbers of alleles expected under the stepwise-mutation model and infinite-allele model deviate from the observed numbers of alleles. The magnitude of this difference is dependent on the sample size, mutation rates (θ-values) and observed gene diversities. Moreover, we show that the number of observed alleles differs among loci with the same observed gene diversity but different mutation rates (θ-values). We propose that a reliable test statistic based on allele excess must determine the confidence interval by computer simulations conditional on the observed gene diversity and θ-values. As the latter are notoriously difficult to obtain for experimental data, we suggest that other statistics, such as ln*RV*, may be better suited to the identification of microsatellite loci subject to selection.

Keywords: neutrality test; microsatellites; allele excess; hitchhiking mapping

1. INTRODUCTION

Over the past decades several theories have been developed to explain the observed levels of variability in natural populations. One interesting subject, which has arisen from this research, is the inference of past selective events from extant natural variability. Several statistical tests have been developed that use sequence variation to distinguish between neutrally evolving genes and selected ones (reviewed in Otto 2000). With the recent progress in high-throughput technology, emphasis is shifting from single-locus studies to complete-genome scans aiming to detect genomic regions that recently acquired a beneficial mutation (Schlötterer 2003). The general idea of such a genome scan is that the fixation of a beneficial mutation in a population also affects sites linked to the target of selection. This phenomenon has been called hitchhiking (Maynard Smith & Haigh 1974; Barton 2000). Hence, screening a large number of markers is expected to identify those linked to a selected site, which therefore deviate from neutral expectations.

Given the high costs of a DNA-sequencing-based genome scan, other more cost-effective genetic markers are required. Microsatellites are highly polymorphic DNA regions distributed over the euchromatic part of the genome in all eukaryotic organisms (Ellegren 2000; Schlötterer 2000). The ease of microsatellite typing in combination with their predominantly neutral evolution renders microsatellites an excellent marker for genome scans (Schlötterer 2004). Nevertheless, in contrast to the analysis of DNA sequences, only a limited number of statistical tests are available to compare observed patterns of microsatellite variability with their neutral expectations.

Microsatellite mutations encompass gains and losses of repeat units (Ellegren 2000; Schlötterer 2000). This stepwise-mutation process was originally studied in the context of protein evolution. The statistical properties of this model have also been used to describe the mutation dynamics of microsatellites. Kimura & Ohta (1975) derived an analytical formula for the expected number of alleles based on the stepwise-mutation model (SMM). The observed number of alleles can be compared with expectations based on the observed gene diversity at this locus. As the observed and expected numbers of alleles should not differ significantly under neutrality, a simple test statistic can be developed. Comparing the observed gene diversity with the gene diversity expected from the number of observed alleles has been used to infer deviations from the stepwise-mutation behaviour of microsatellites (Shriver *et al.* 1993; Estoup *et al.* 1995) and neutrality of microsatellite variability (Michalakis & Veuille 1996) and to identify single loci deviating from neutral expectations (Payseur *et al.* 2002; Vigouroux *et al.* 2002*b*). As the same discrepancy is tested in these scenarios, i.e. observed and expected gene diversities conditional on the observed number of alleles or the observed and expected numbers of alleles conditional on the observed gene diversity, we will refer to these statistical tests collectively as the allele-excess test statistic.

The analytical formulae were tested for low mutation rates only (Kimura & Ohta 1975). Recent computer simulations have demonstrated that the expected number of alleles is underestimated for loci with a high mutation rate (Shriver *et al.* 1993). Therefore, a recently developed test that compares observed and expected gene diversities relies on computer simulations to determine the expected gene diversity (Cornuet & Luikart 1996). In this report we focus on the suitability of the allele-excess test statistic for the inference of selection at individual microsatellite loci.

2. MATERIAL AND METHODS

We used a commonly employed coalescent-based computer simulation algorithm (Hudson 1990), which has been modified

*Author for correspondence (christian.schloetterer@vu-wien.ac.at).

DOI 10.1098/rspb.2003.2662

to account for the stepwise-mutation behaviour of microsatellites. Rather than counting the number of mutations occurring on a branch, our simulations traced the allele length of a microsatellite locus. The number of mutations occurring on a branch was converted into microsatellite mutations by adding or removing (with equal probability) one repeat unit for each mutation. The accuracy of the code was checked by comparing the observed variance in repeat number with its expectation, $E(V)$, $(E(V) = \theta/2)$. Gene diversities, H, were calculated as

$$H = \frac{n}{n-1}\left(1 - \sum_{i=1}^{m} x_i^2\right),$$

where m is the number of alleles, n is the number of analysed chromosomes and x is the allele frequency. Calculated gene diversities were verified by using the simulated allele frequencies as input in the MSA software, which was independently written to compute microsatellite-specific statistics (Dieringer & Schlötterer 2003). Finally, we tested the code with a different microsatellite-evolution program, which uses a different algorithm to generate random numbers and Poisson deviates (kindly provided by T. Wiehe).

Unless otherwise noted, θ-values were drawn from a uniform distribution between 0.1 and 10.1 to account for heterogeneity in microsatellite mutation rates. We simulated 30 000 unlinked loci for each combination of parameters and a sample size of 100 chromosomes.

We also performed computer simulations accounting for the observed distribution of microsatellite variability in natural populations. Previous studies have demonstrated that the natural logarithm of the observed variance in a repeat number follows a normal distribution (Goldstein *et al.* 1996; Harr *et al.* 1998). Therefore, we used the mean (1.96) and standard deviation (1.28) of the natural logarithm of V observed in African *Drosophila melanogaster* populations (Caracristi & Schlötterer 2003) to describe the normal distribution from which we sampled the log $\theta/2$-values for our computer simulations.

The expected number of alleles under the SMM was calculated as described by Kimura & Ohta (1975):

$$n_{\text{expected}} = \frac{\theta + \beta}{\beta}\left\{1 - \prod_{i=0}^{2N-1}\left(\frac{i+\theta}{i+\theta+\beta}\right)\right\}, \tag{2.1}$$

with

$$\theta = 4N_e\mu = \left(\frac{1}{H_o^2} - 1\right)\frac{1}{2} \tag{2.2}$$

and

$$\beta = \frac{\theta + 1 - \sqrt{1 + 8N_e\mu}}{\sqrt{1 + 8N_e\mu} - 1} = \frac{\theta + 1 - \dfrac{1}{H_o}}{\dfrac{1}{H_o} - 1} = \frac{H_o\theta + H_o - 1}{1 - H_o}. \tag{2.3}$$

The expected number of alleles under the infinite-allele model (IAM) was calculated as described by Watterson (1975):

$$n_{\text{expected}} = \sum_{i=1}^{2N}\frac{\theta}{\theta + i - 1}, \tag{2.4}$$

with

$$\theta = \frac{1 - H_o}{H_o}. \tag{2.5}$$

N_e is the effective diploid population size, N is the number of diploid individuals in the analysed sample, H_o is the expected homozygosity and μ is the microsatellite mutation rate.

Allele excess was determined as

$$\text{AE} = \frac{n_{\text{observed}} - n_{\text{expected}}}{n_{\text{expected}}}. \tag{2.6}$$

3. RESULTS

We simulated 30 000 microsatellite loci using a broad range of sample sizes and θ-values. For each dataset the mean allele excess was determined based on the SMM and the IAM. While for most simulations the SMM resulted in allele excess, the IAM indicated an allele deficiency (table 1). The closest fit to the expectation for both models was obtained for very low θ-values $(\theta = 0.05)$. The SMM also provided a good fit to the expectations in the simulation based on small sample sizes $(n = 10)$. Interestingly, the mean allele excess was strongly influenced by the sample size. For the SMM a larger sample size resulted in more pronounced allele excess, while a more extreme allele deficiency was observed at large sample sizes under the IAM. The same trend was observed when no correction for sample size was made to the gene-diversity estimate (data not shown). We also found that θ-values were linked to allele excess. An increase in θ resulted in a higher allele excess under the SMM and a larger allele deficiency under the IAM. Deviations from the strict SMM that allowed for larger changes in repeat number (two-phase model; Di Rienzo *et al.* 1994) increased the allele excess under the SMM, but resulted in a less pronounced allele deficiency under the IAM (table 1).

As some tests of neutrality based on the excess of alleles focus on individual loci, we were interested in the distribution of allele excess over the range of the simulated data. To investigate this, we grouped data simulated from a broad range of θ-values into different (observed) gene-diversity classes and determined the allele excess for each of the classes separately. For an unbiased test statistic the observed allele excess should be independent of the observed gene diversity. As expected, no allele excess was observed for monomorphic loci (figure 1). Very strong allele excess was observed for loci with low levels of observed gene diversity $(0 < H < 0.1$; figure 1). Interestingly, both the IAM and SMM resulted in almost the same very strong allele excess. Only for larger gene diversities did the difference between the two mutation models become apparent (figure 1). Further computer simulations based on different sample sizes consistently indicated that the lowest-gene-diversity class (excluding $H = 0$) had the most extreme allele excess (figure 2*a*,*b*). The same phenomenon was observed when computer simulations were performed with fixed θ-values. Irrespective of the θ-value used, the lowest-gene-diversity class had the most pronounced allele excess (figure 3*a*,*b*). These results clearly indicate that the average allele excess and its confidence interval are not well suited to determining the significance of allele excess, as the mean allele excess differs substantially among different classes of observed gene diversity. One further problem of the allele excess becomes apparent when comparing different θ-values (figure 3). As expected, simulations based on

Table 1. Comparison of the allele excess based on the SMM and IAM using simulated microsatellite data. (Standard deviations of all means are given in brackets.)

θ sample size (N)	0.1–10.1 10	0.1–10.1 100	0.1–10.1 1000	0.05 100	0.5 100	5.0 100	50.0 100	0.1–10.1[a] 100
mean θ	5.116 (±2.882)	5.097 (±2.884)	5.089 (±2.901)	0.050	0.500	5.000	50.000	5.092 (±2.881)
mean V	2.583 (±4.174)	2.532 (±3.670)	2.530 (±3.778)	0.025 (±0.072)	0.250 (±0.354)	2.491 (±2.863)	25.077 (±29.271)	5.276 (±7.913)
mean H	0.650 (±0.214)	0.648 (±0.186)	0.648 (±0.184)	0.047 (±0.120)	0.293 (±0.207)	0.699 (±0.097)	0.901 (±0.028)	0.684 (±0.188)
mean number of observed alleles	3.723 (±1.337)	5.753 (±2.203)	6.556 (±2.473)	1.238 (±0.458)	2.477 (±0.826)	5.998 (±1.489)	15.555 (±3.371)	7.176 (±2.830)
expected number of alleles SMM	3.904 (±1.403)	5.218 (±1.776)	5.356 (±1.661)	1.237 (±0.607)	2.511 (±1.104)	5.462 (±1.205)	12.710 (±2.867)	5.823 (±2.108)
expected number of alleles IAM	4.310 (±1.520)	9.418 (±4.089)	15.021 (±7.239)	1.338 (±0.934)	3.367 (±2.033)	9.981 (±2.914)	23.842 (±4.312)	10.789 (±4.744)
(o − e)/e SMM[b]	−0.031 (±0.163)	0.118 (±0.265)	0.232 (±0.290)	0.042 (±0.233)	0.104 (±0.416)	0.113 (±0.227)	0.237 (±0.177)	0.249 (±0.302)
(o − e)/e IAM[b]	−0.118 (±0.159)	−0.324 (±0.262)	−0.478 (±0.298)	0.023 (±0.255)	−0.047 (±0.480)	−0.371 (±0.172)	−0.346 (±0.088)	−0.269 (±0.275)

[a] In this column 30% of the microsatellite mutations encompassed more than a single repeat unit. The upper boundary for a change in repeat number was fixed at three repeats.

[b] (o − e)/e, difference between observed and expected expressed as a fraction of the expected number of alleles.

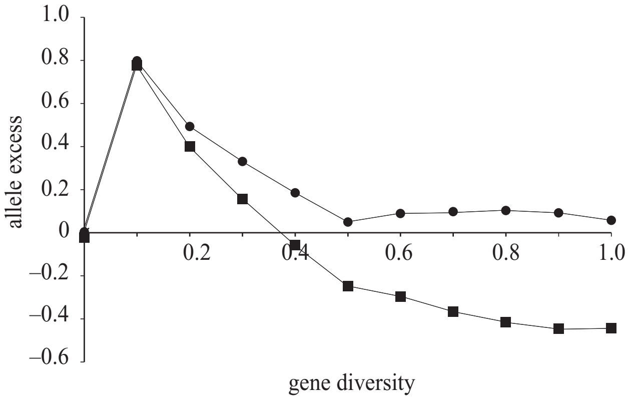

Figure 1. Dependence of allele excess on observed gene diversity (squares, IAM; circles, SMM). Gene diversities obtained from neutral coalescent simulations were grouped into 11 bins and for each bin the mean allele excess is shown. Parameters for the coalescent simulations were: $\theta = 0.1–10.1$, $N = 100$ chromosomes.

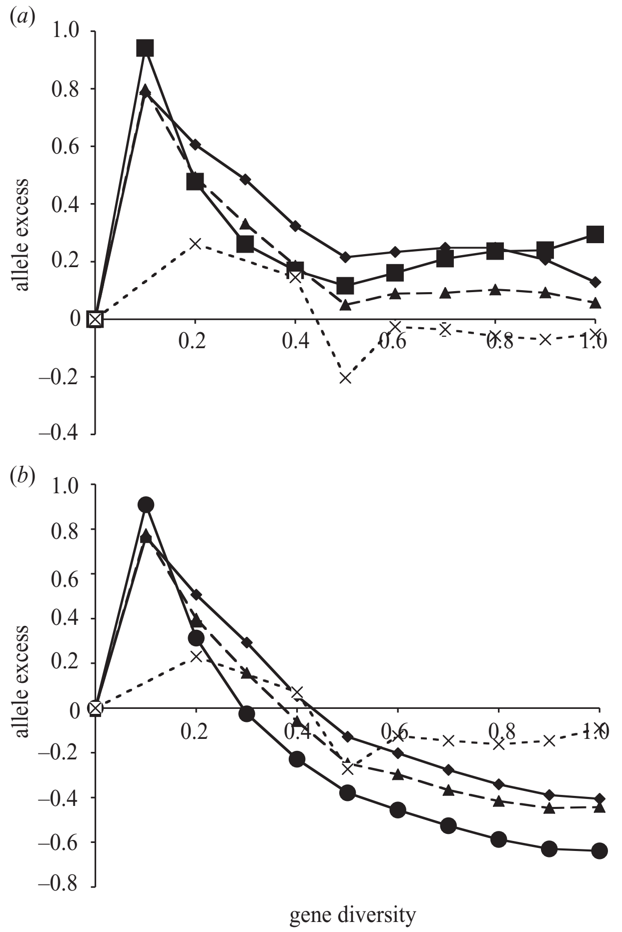

Figure 2. Allele excess and sample size. (*a*) SMM (diamonds, $N = 100$, two-phase; squares, $N = 1000$; triangles, $N = 100$; crosses, $N = 10$). (*b*) IAM (diamonds, $N = 100$, two-phase; circles, $N = 1000$; triangles, $N = 100$; crosses, $N = 10$). Gene diversities obtained from neutral coalescent simulations were grouped into 11 bins and for each bin the mean allele excess is shown. For all simulations θ was drawn from a uniform distribution between 0.1 and 10.1.

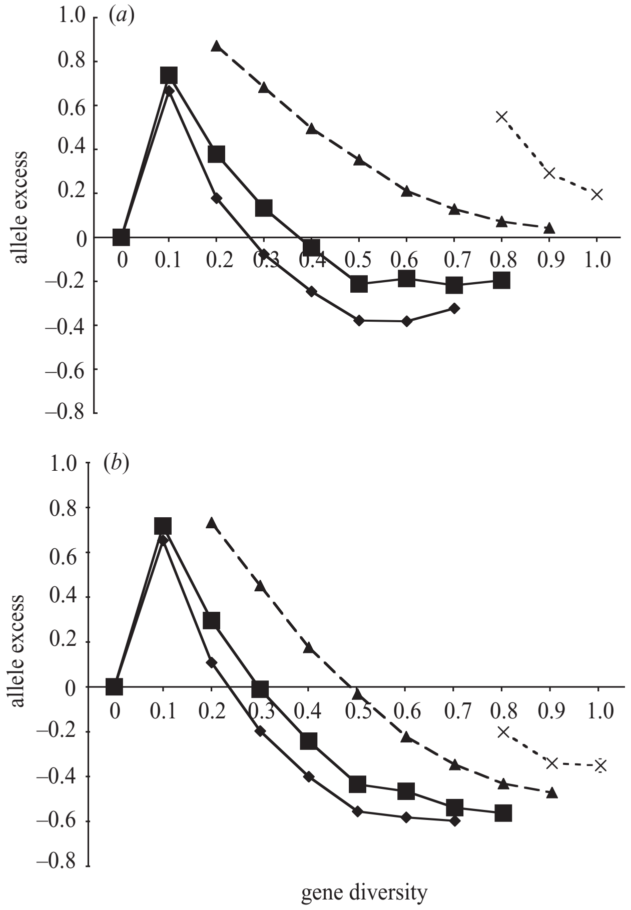

Figure 3. Allele excess and θ. (*a*) SMM (diamonds, $\theta = 0.05$; squares, $\theta = 0.5$; triangles, $\theta = 5$; crosses, $\theta = 50$). (*b*) IAM (diamonds, $\theta = 0.05$; squares, $\theta = 0.5$; triangles, $\theta = 5$; crosses, $\theta = 50$). Gene diversities obtained from neutral coalescent simulations were grouped into 11 bins and for each bin the mean allele excess is shown. All simulations are based on a sample size of 100 chromosomes.

different θ-values generate overlapping distributions of gene diversities. As a consequence, loci with gene diversities between 0.4 and 0.5 showed a mean allele deficiency (AE = −0.21, SMM) when they were simulated with θ-values of 0.5. For simulations using an θ-value of 5.0, loci with a gene diversity between 0.4 and 0.5 showed an allele excess (AE = 0.35, SMM). This difference is highly significant ($p < 0.0001$, Mann–Whitney *U*-test), indicating that, despite a similar gene diversity, the number of alleles at a given locus is determined by its mutation rate.

As the analytical formula by Kimura & Ohta (1975) underestimates the expected number of alleles for loci with high mutation rates, the difference in allele excess could also result from this bias. Therefore, we compared the numbers of observed alleles for different θ-values (table 2), but the same trend could be recognized. Depending on the θ-value used for the computer simulation, we detected substantial differences in the mean number of alleles observed within a gene-diversity class.

So far, we have considered only θ-values drawn from a uniform distribution. As the natural logarithm of the variance in repeat number follows a normal distribution (Goldstein *et al.* 1996; Harr *et al.* 1998), we also performed computer simulations using log θ-values drawn from a normal distribution where the mean and standard deviation were estimated from an African *D. melanogaster* population (Caracristi & Schlötterer 2003). As expected, we observed that simulation runs with low gene diversity showed a pronounced allele excess (figure 4). We also examined the allele excess when $\theta = 4N_e\mu$ was determined by gene diversity (equation (2.2)) or by the variance in repeat number ($\theta = 2V$). Both estimators showed the pronounced surplus of expected alleles for small gene diversities. For larger gene diversities, however, the variance-based estimator showed a more pronounced allele excess than did the gene-diversity based one (figure 4).

4. DISCUSSION

Our simulations indicate that the mean numbers of observed alleles for small sample sizes and low θ-values are very similar to those predicted under the SMM (Kimura & Ohta 1975). For larger sample sizes and higher θ-values (in the range typical for microsatellites) we found a large discrepancy between the observed number of alleles and the expectation based on the analytical formulae (Kimura & Ohta 1975). Our observation is in qualitative agreement with that of a previous simulation study (Shriver *et al.* 1993).

Currently, allele excess is widely used for the identification of loci that are affected by natural selection (Estoup *et al.* 1995; Payseur *et al.* 2002; Vigouroux *et al.* 2002*b*). The challenge for such tests, particularly for genome scans, is that several evolutionary forces are influencing variability in natural populations. For example, low gene diversity at a microsatellite locus may have different causes: (i) the microsatellite may have a low mutation rate, resulting in a lower expected gene diversity than for loci with higher mutation rates; (ii) even loci with high mutation rates could have low levels of gene diversity if the sampled alleles share a common ancestor in the past; and (iii) hitchhiking: if a microsatellite locus is closely linked to a genomic region that recently experienced a selective sweep, this microsatellite will have lower levels of variability. The task of any neutrality test is to interpret observed variation so as to distinguish the two neutral scenarios ((i) and (ii)) from the selection hypothesis. The underlying idea is that under neutrality the observed number of alleles should be consistent with the observed gene diversity. If more alleles are observed than expected from the observed gene diversity, this is regarded as evidence for selection. Our computer simulations indicated two possible complications in using this approach. First, the analytical formula underestimates the expected number of alleles, leading to allele excess. This problem could be solved by using computer simulations to predict the expected number of alleles conditional on the observed gene diversity (or alternatively to determine the expected gene diversity conditional on the observed number of alleles). Second, the average number of alleles at a locus with a given gene diversity depends on their mutation rates (θ-values). Note that this observation is independent of the analytical formula used to determine the expected number of alleles. Thus computer simulations conditioning on the observed gene diversity are not well suited to addressing this discrepancy. An unbiased test statistic would require computer simulations conditional on the observed gene diversity and the θ-values (a joint estimator

Table 2. Observed numbers of alleles in different classes of gene diversity ($N = 100$).

gene diversity	$\theta = 0.5$	$\theta = 5$	$\theta = 50$
0	1 (±0)	—	—
> 0–0.1	2.19 (±0.40)	—	—
> 0.1–0.2	2.43 (±0.54)	3.50 (±0.67)	—
> 0.2–0.3	2.55 (±0.59)	3.90 (±0.94)	—
> 0.3–0.4	2.62 (±0.64)	4.21 (±0.99)	—
> 0.4–0.5	2.60 (±0.67)	4.49 (±1.04)	—
> 0.5–0.6	3.08 (±0.64)	4.76 (±1.05)	—
> 0.6–0.7	3.57 (±0.65)	5.34 (±1.10)	—
> 0.7–0.8	4.39 (±0.55)	6.34 (±1.18)	10.18 (±1.94)
> 0.8–0.9	—	7.98 (±1.26)	13.13 (±2.28)
> 0.9–1	—	—	17.42 (±2.82)

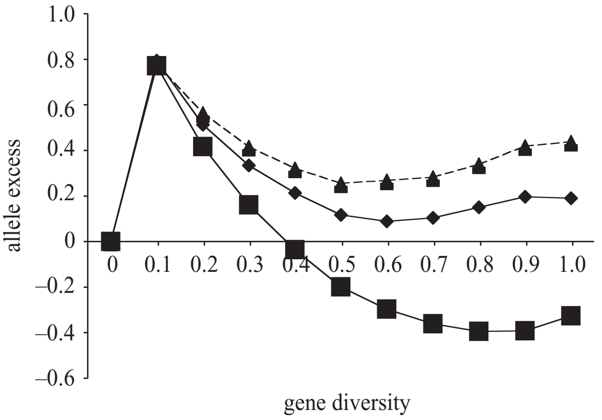

Figure 4. Allele excess for computer simulations based on a normal distribution of log θ-values. Gene diversities obtained from neutral coalescent simulations were grouped into 11 bins and for each bin the mean allele excess is shown. All simulations are based on a sample size of 100 chromosomes. The expected number of alleles is based either on the observed gene diversities (diamonds, SMM; squares, IAM) or on the observed variance in repeat number (triangles, SMM).

of the microsatellite mutation rate and the effective population size). As mutation rates differ substantially among loci (Di Rienzo *et al.* 1994; Harr *et al.* 1998; Vigouroux *et al.* 2002*a*) in most experimental surveys of microsatellite variation, the required θ-values are not available. Consequently, it is extremely difficult to obtain an unbiased significance level for the allele-excess test statistic.

The outcome of the complex behaviour of the allele-excess test statistic is indicated in figure 1. Our computer simulations show that even under neutrality the mean allele excess was greatly elevated for loci with low observed gene diversity. Consistent with this observation, those loci that were identified by the allele-excess test statistic as significant outliers had low gene diversities (Vigouroux *et al.* 2002*b*). Similar results were obtained in a genome scan in *Drosophila*, which found that loci with a low gene diversity were more likely to have an excess of alleles (Kauer *et al.* 2003).

5. CONCLUSION

We demonstrated a strong dependence of the allele-excess test statistic on both θ and the gene diversity of each microsatellite locus. Given the difficulty in obtaining reliable locus-specific θ estimates, we suggest that results obtained with the allele-excess test statistic should be treated with caution. An alternative to the use of allele excess is a recently suggested statistic (lnRV) that also uses microsatellite polymorphism for the inference of selective sweeps (Schlötterer 2002). Rather than contrasting observed and expected allele numbers, this test compares levels of variability for each locus in two populations. By calculating the ratio of the observed variances in repeat number, this statistic has an identical expectation for all loci, independent of their θ-values.

Finally, we note that similar problems will be encountered for DNA sequence data. Neutrality tests attempting to infer non-neutral evolution are based on θ-values estimated from polymorphism data. In contrast to the problem with microsatellites, this problem could be alleviated by the use of mutation-rate estimates from between-species divergence in combination with reliable estimates of population size.

We are grateful to T. Wiehe for sharing his code for the simulation of microsatellite data. R. Bürger, B. Harr and G. Muir provided helpful comments on the manuscript. This work has been supported by grants from the Fonds zur Förderung der wissenschaftlichen Forschung (FWF) and an EMBO young investigator award to C.S.

REFERENCES

Barton, N. H. 2000 Genetic hitchhiking. *Phil. Trans. R. Soc. Lond.* B **355**, 1553–1562. (DOI 10.1098/rstb.2000.0716.)

Caracristi, G. & Schlötterer, C. 2003 Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. *Mol. Biol. Evol.* **20**, 792–799.

Cornuet, J. M. & Luikart, G. 1996 Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* **144**, 2001–2014.

Dieringer, D. & Schlötterer, C. 2003 Microsatellite analyzer (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Notes* **3**, 167–169.

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M. & Freimer, N. B. 1994 Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl Acad. Sci. USA* **91**, 3166–3170.

Ellegren, H. 2000 Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.* **16**, 551–558.

Estoup, A., Garnery, L., Solignac, M. & Cornuet, J.-M. 1995 Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* **140**, 679–695.

Goldstein, D. B., Zhivotovsky, L. A., Nayar, K., Ruiz Lineares, A., Cavalli-Sforza, L. L. & Feldman, M. W. 1996 Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. *Mol. Biol. Evol.* **13**, 1213–1218.

Harr, B., Zangerl, B., Brem, G. & Schlötterer, C. 1998 Conservation of locus specific microsatellite variability across species: a comparison of two *Drosophila* sibling species *D. melanogaster* and *D. simulans*. *Mol. Biol. Evol.* **15**, 176–184.

Hudson, R. R. 1990 Gene geneologies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–44.

Kauer, M. O., Dieringer, D. & Schlötterer, C. 2003 A microsite variability screen for positive selection associated with the 'out of Africa' habitat expansion of *Drosophila melanogaster*. *Genetics* **165**, 1137–1148.

Kimura, M. & Ohta, T. 1975 Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proc. Natl Acad. Sci. USA* **72**, 2761–2764.

Maynard Smith, J. & Haigh, J. 1974 The hitch-hiking effect of a favorable gene. *Genet. Res.* **23**, 23–35.

Michalakis, Y. & Veuille, M. 1996 Length variation of CAG/CAA trinucleotide repeats in natural populations of *Drosophila melanogaster* and its relation to the recombination rate. *Genetics* **143**, 1713–1725.

Otto, S. P. 2000 Detecting the form of selection from DNA sequence data. *Trends Genet.* **16**, 526–529.

Payseur, B. A., Cutter, A. D. & Nachman, M. W. 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* **19**, 1143–1153.

Schlötterer, C. 2000 Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**, 365–371.

Schlötterer, C. 2002 A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* **160**, 753–763.

Schlötterer, C. 2003 Hitchhiking mapping: functional genomics from the population genetics perspective. *Trends Genet.* **19**, 32–38.

Schlötterer, C. 2004 The evolution of molecular markers—just a matter of fashion? *Nat. Rev. Genet.* **5**, 63–69.

Shriver, M. D., Jin, L., Chakraborty, R. & Boerwinkle, E. 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics* **134**, 983–993.

Vigouroux, Y., Jaqueth, J. S., Matsuoka, Y., Smith, O. S., Beavis, W. D., Smith, J. S. & Doebley, J. 2002*a* Rate and pattern of mutation at microsatellite loci in maize. *Mol. Biol. Evol.* **19**, 1251–1260.

Vigouroux, Y., McMullen, M., Hittinger, C. T., Houchins, K., Schulz, L., Kresovich, S., Matsuoka, Y. & Doebley, J. 2002*b* Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proc. Natl Acad. Sci. USA* **99**, 9650–9655.

Watterson, G. A. 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276.