

## PROGRAM NOTE

**MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data**

COCK VAN OOSTERHOUT,\* WILLIAM F. HUTCHINSON,\* DEREK P. M. WILLS† and PETER SHIPLEY\*†

\*Department of Biological Sciences and †Department of Computer Sciences, University of Hull, Hull, HU6 7RX, UK

**Abstract**

DNA degradation, low DNA concentrations and primer-site mutations may result in the incorrect assignment of microsatellite genotypes, potentially biasing population genetic analyses. MICRO-CHECKER is WINDOWS®-based software that tests the genotyping of microsatellites from diploid populations. The program aids identification of genotyping errors due to nonamplified alleles (null alleles), short allele dominance (large allele dropout) and the scoring of stutter peaks, and also detects typographic errors. MICRO-CHECKER estimates the frequency of null alleles and, importantly, can adjust the allele and genotype frequencies of the amplified alleles, permitting their use in further population genetic analysis. MICRO-CHECKER can be freely downloaded from <http://www.microchecker.hull.ac.uk/>.

*Keywords:* MICRO-CHECKER, microsatellite DNA, stuttering, null alleles, PCR slippage, short allele dominance

*Received 3 December 2003; revision received 20 February 2004; accepted 22 March 2004*

Microsatellites are currently the most widely used genetic markers in evolutionary and ecological studies. Their popularity stems from their near ubiquity, generally high level of polymorphism, codominance, and relative ease of screening once isolated. Furthermore, only a few DNA copies suffice for PCR-amplification, which has proved invaluable in conservation studies of contemporary populations, as well as in long-term temporal genetic analyses utilizing ancient DNA (Hutchinson *et al.* 2003). Paradoxically, these potential advantages can also increase the incidence of genotyping errors, which has received insufficient attention to date.

Genotyping errors can be caused by low template DNA concentrations (Wandeler *et al.* 2003), which may result in an allele failing to amplify due to stochastic sampling error (allelic dropout, Miller & Waits 2003). Genotyping errors can also occur due to the preferential amplification of small alleles (i.e. large allele dropout or short allele dominance, Wattier *et al.* 1998), where the larger allele specifically fails to amplify. Furthermore, slippage during PCR-amplification can produce additional stutter products that differ from the original template by multiples of the repeat unit length

(Shinde *et al.* 2003). Such stutters are often common in dinucleotide loci, making it difficult to discriminate between homozygotes and heterozygotes. Finally, where mutations occur at primer sites, certain alleles may not be amplified (null alleles) resulting in false homozygotes (Shaw *et al.* 1999).

Such genotyping errors can cause deviations from Hardy–Weinberg proportions, in particular a heterozygote deficiency (Shaw *et al.* 1999), potentially biasing population genetic analyses. These deviations are often very similar to those caused by inbreeding, assortative mating or Wahlund effects. Nevertheless, short allele dominance, stuttering and null alleles yield their own specific allelic ‘signature’ (i.e. deficiencies and excesses of particular genotypes), as opposed to allelic dropout, which is assumed to be largely independent of allele size (Miller *et al.* 2002; Miller & Waits 2003). It is therefore possible to discriminate between deviations due to nonpanmixia and those caused by the various genotyping errors.

There are currently a variety of programs that help with the detection of genotyping errors. GIMLET (Valière 2002) can identify false homozygotes and false alleles, but this program requires repeated genotyping of each sample. Similarly, replication of genotyping is also required for the maximum likelihood approach of Miller *et al.* (2002). Software such as PEDMANAGER (Ewen *et al.* 2000) applies

Correspondence: Cock van Oosterhout. Fax: + 44 (0)1482 465458; E-mail: C.van-Oosterhout@hull.ac.uk

Mendelian-inheritance-error checking, which requires family pedigree records. CERVUS version 2.0 (Marshall *et al.* 1998) establishes the presence of null alleles by analysing deviations from Hardy–Weinberg equilibrium using a chi-square goodness-of-fit test, and uses an iterative algorithm based on the difference between observed and expected frequency of homozygotes to estimate the null allele frequency (Summers & Amos 1997). However, CERVUS 2.0 is unable to discriminate null alleles from other genotyping errors and does not provide alternative methods for estimating null alleles (see below).

Chakraborty *et al.* (1992) and Brookfield (1996) discuss several methods to estimate the frequency of null alleles ( $r$  in their terminology) from an apparent heterozygote deficiency. The choice of method varies depending on whether some samples failed to amplify, and on whether these nonamplified samples represent null homozygotes (see Brookfield 1996). Furthermore, these methods assume that the heterozygote deficiency is caused by null alleles and not by other genotyping errors or deviations from panmixia.

We have developed a new software tool, MICRO-CHECKER, which can help identify genotyping errors due to null alleles,

short allele dominance (large allele dropout) and scoring errors due to stuttering. Where multi-locus genotypes are available, it can discriminate between inbreeding and Wahlund effects, and Hardy–Weinberg deviations caused by null alleles. Furthermore, it can identify possible typographic errors. Finally, if null alleles are present, MICRO-CHECKER can estimate the null allele frequency, and adjust the observed allele and genotype frequencies accordingly. These adjusted allele frequencies can be used subsequently for further population genetic analysis. Below, we provide a brief description of the methods implemented in MICRO-CHECKER.

The program accepts a variety of input formats for microsatellite data including GENEPOP (Raymond & Rousset 1995), Microsoft Excel and text file formats (ACSI). Following a check for potential typographical errors, the program constructs random genotypes by randomizing the observed alleles for each locus within samples. The observed genotypes are then compared with the distribution of randomized genotypes, and results displayed in two graphs (Fig. 1). The left-hand graph shows the frequency of the allele-specific homozygotes (which we will refer to

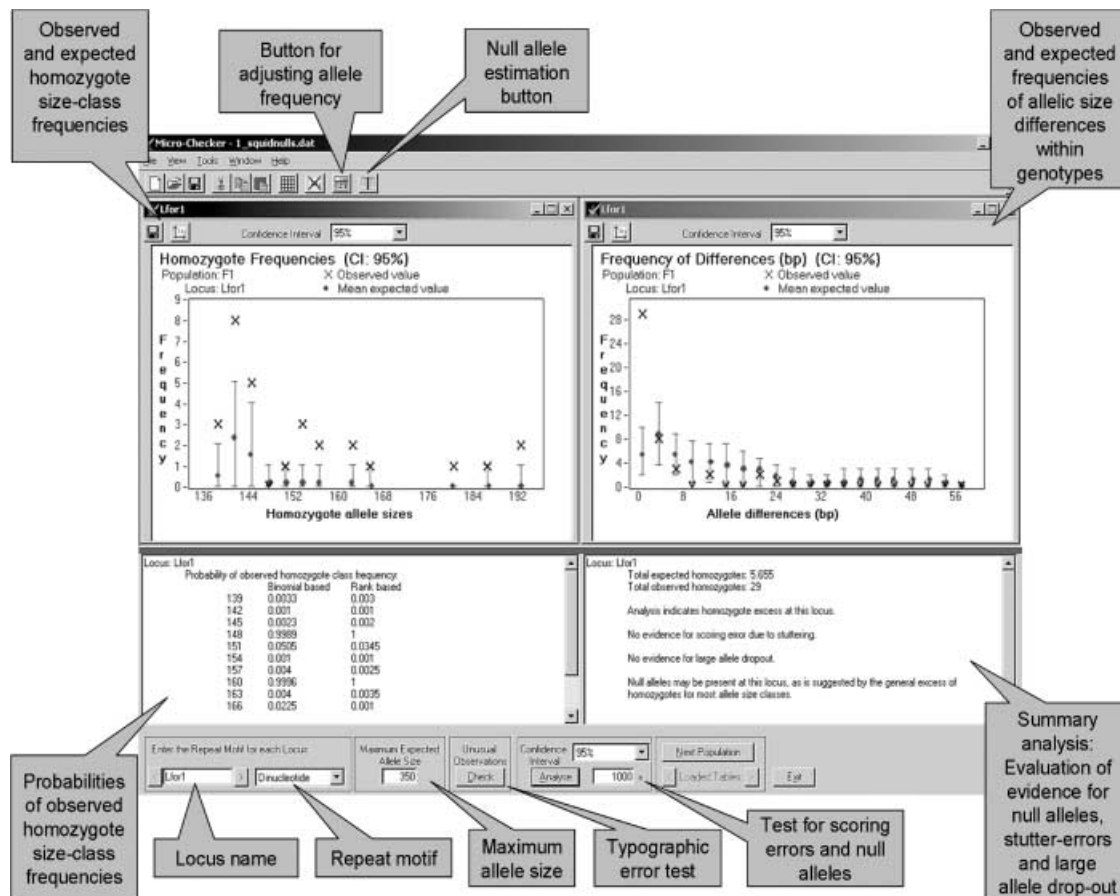


Fig. 1 Output of the MICRO-CHECKER program including graphs, probabilities for all observed homozygote-classes, and a concise summary of the analysis.

as homozygote size classes). The right-hand graph shows the frequency of genotypes categorized by the difference in base pairs between both alleles (the so-called heterozygote size classes). Probabilities for the observed number of homozygotes within each homozygote size class (lower left panel in Fig. 1) are calculated using a cumulative binomial distribution (Weir 1996). The  $P$ -values are also calculated by ranking the observed frequency in the distribution of randomized genotypes. Fisher's combined probability test is then calculated using the  $P$ -values of all homozygote-classes. Finally, significant deviations from Hardy-Weinberg proportions are identified, and it is indicated whether these are consistent with null alleles, short allele dominance, or scoring errors associated with stuttering (lower right panel in Fig. 1).

If there is an overall homozygote excess that, crucially, is not homogeneously distributed across all homozygote-classes, the pattern may not be consistent with null alleles, and MICRO-CHECKER will warn of possible genotyping errors or deviations from panmixia. The mis-scoring of stutters is suggested when there is a deficiency of heterozygotes with alleles differing in size by a single repeat, and a relative excess of large homozygotes (because stuttering is more severe for larger alleles). Short allele dominance is suggested when an excess of homozygotes is biased towards either extreme of the allele size-distribution. In addition, when there is a general homozygote excess and the allelic range exceeds 150 base pairs, MICRO-CHECKER will caution for possible short allele dominance. Examples of genotyping errors and suggestions on how to correct for them are given in the Help Files and User Manual.

MICRO-CHECKER indicates the presence of null alleles if the combined probability test shows there is an overall significant excess of homozygotes, and when this is evenly distributed across the homozygote-classes. However, if all loci show an excess of homozygotes, the program warns that the population might not be panmictic. The program will calculate the null allele frequency ( $r$ ), using the methods described by Chakraborty *et al.* (1992) and Brookfield (1996), if null alleles are indicated.

If the observed deviations in genotype frequencies are consistent with the segregation of a null allele, MICRO-CHECKER can also adjust the frequencies of the amplified alleles to account for the downward bias resulting from the null allele, and thereby assumes that the population is in Hardy-Weinberg equilibrium. If  $p$  is the proportion of  $A$  alleles in the population,  $r$  the proportion of null alleles (0),  $n_1$  the proportion of two-banded individuals (i.e. the real  $A$  heterozygotes minus  $A0$ ),  $n_2$  the proportion of one-banded individuals (i.e.  $AA$  homozygotes plus  $A0$  heterozygotes), then the expected proportion of genotypes in the population equals:

$$n_1 = A^* - A0 = 2p(1 - p) - 2pr, \quad (1a)$$

$$n_2 = AA + A0 = p^2 + 2pr. \quad (1b)$$

Solving these quadratic equations gives two estimates for the allele frequency  $p$ , based on the number of one and two-banded individuals, respectively:

$$p = \frac{1}{2}(-2r + [(2r^2 + 4n_2)]^{1/2}). \quad (2a)$$

$$p = \frac{1}{4}(-2r - 1) + [(2r - 2)^2 - 8n_1]^{1/2}. \quad (2b)$$

The program calculates the adjusted frequencies of all amplified alleles, along with the frequencies of the genotypes (adjusted using eqn 1b). These are no longer multi-locus genotypes, but both allele and genotype frequencies can be used for further population genetic analysis.

MICRO-CHECKER, including the User Manual, Help and Example files can be freely downloaded from <http://www.microchecker.hull.ac.uk/>.

## Acknowledgements

We wish to thank Gary R. Carvalho, Greg Adcock, David Weetman, Africa Gomez, Helen Wilcock, Bernd Hänfling, Richard Case, Natasha Mesquita, Eugenia D'Amato, Rob van Treuren, Lars-Erik Holm and Adam G. Jones for providing test data and/or helpful comments during the development of the software program. The work was supported by the Natural Environment Research Council, UK (NERC Fellowship NER/I/S/2000/00885 to CVO), and a University of Hull Research Support grant.

## References

- Brookfield JFY (1996) A simple new method for estimating null allele frequency from heterozygote deficiency. *Molecular Ecology*, **5**, 453–455.
- Chakraborty R, De Andrade M, Daiger SP, Budowle B (1992) Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. *Annals of Human Genetics*, **56**, 45–57.
- Ewen KR, Bahlo M, Treloar SA, Levinson DF, Mowry B, Barlow JW, Foote SJ (2000) Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics*, **67**, 727–736.
- Hutchinson WF, Van Oosterhout C, Rogers SI, Carvalho GR (2003) Temporal analysis of archived samples indicates marked genetic changes in declining North Sea cod (*Gadus morhua*). *Proceedings of the Royal Society London B*, **270**, 2125–2132.
- Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **7**, 639–655.
- Miller CR, Waits LP (2003) The history of effective population size and genetic diversity in the Yellowstone grizzly (*Ursus arctos*): Implications for conservation. *Proceedings of the National Academy of Sciences USA*, **100**, 4334–4339.
- Miller CR, Joyce P, Waits LP (2002) Assessing allelic dropout and genotyping reliability using maximum likelihood. *Genetics*, **160**, 357–366.
- Raymond M, Rousset F (1995) GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

- Shaw PW, Pierce GJ, Boyle PR (1999) Subtle population structuring within a highly vagile marine invertebrate, the veined squid *Loligo forbesi*, demonstrated with microsatellite DNA markers. *Molecular Ecology*, **8**, 407–417.
- Shinde D, Lai YL, Sun FZ, Arnheim N (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)(n) and (A/T)(n) microsatellites. *Nucleic Acid Research*, **31**, 974–980.
- Summers K, Amos W (1997) Behavioral, ecological and molecular genetic analyses of reproductive strategies in the Amazonian dart-poison frog, *Dendrobates ventrimaculatus*. *Behavioral Ecology*, **8**, 260–267.
- Valière N (2002) GIMLET: a computer program for analysing genetic individual identification data. *Molecular Ecology*, **2**, 377–379.
- Wandeler P, Smith S, Morin PA, Pettifor RA, Funk SM (2003) Patterns of nuclear DNA degeneration over time — a case study in historic teeth samples. *Molecular Ecology*, **12**, 1087–1093.
- Wattier R, Engel CR, Saumitou-Laprade P, Valero M (1998) Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in *Gracilaria gracilis* (Rhodophyta). *Molecular Ecology*, **7**, 1569–1573.
- Weir BS (1996) *Genetic Data Analysis II*. Sinauer Associates, Sunderland, Massachusetts.