Molecular Ecology Notes (2007) **7**, 10–14 doi: 10.1111/j.1471-8286.2006.01560.x

TECHNICAL ARTICLE

Automated binning of microsatellite alleles: problems and solutions

W. AMOS,* J. I. HOFFMAN,* A. FRODSHAM,† L. ZHANG,† S. BEST† and A. V. S. HILL†
**Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, UK, †The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK*

Abstract

As genotyping methods move ever closer to full automation, care must be taken to ensure that there is no equivalent rise in allele-calling error rates. One clear source of error lies with how raw allele lengths are converted into allele classes, a process referred to as binning. Standard automated approaches usually assume collinearity between expected and measured fragment length. Unfortunately, such collinearity is often only approximate, with the consequence that alleles do not conform to a perfect 2-, 3- or 4-base-pair periodicity. To account for these problems, we introduce a method that allows repeat units to be fractionally shorter or longer than their theoretical value. Tested on a large human data set, our algorithm performs well over a wide range of dinucleotide repeat loci. The size of the problem caused by sticking to whole numbers of bases is indicated by the fact that the effective repeat length was within 5% of the assumed length only 68.3% of the time.

Keywords: ABI, allele calling, fragment size, GENOTYPER, genotyping error.

Received 20 April 2006; revision accepted 21 August 2006

Despite increasing interest in single nucleotide polymorphisms, microsatellites remain arguably the most important class of markers for use in population genetic studies (Schlotterer 2004). Since microsatellites were first used as markers around 1990, there has been a steady switch in the way alleles are visualized, from radioactive incorporation to fluorescent dye technology. In theory, fluorescent methods allow much greater levels of automation, with internal size markers aiding size estimation and various algorithms assigning fragments to allele classes. However, the reality appears to be that progression from raw fragment lengths to allele sizes remains an important challenge (Ginot *et al*. 1996; Ewen *et al*. 2000; Weeks *et al*. 2002).

Several strategies have been developed for allele assignment. Where allele size distributions are already well documented, for example in European human populations, any particular fragment can be compared against a database of expected lengths and assigned to the closest. This seems to work well, but becomes prone to error when new alleles are found, and is of course inappropriate for most nonhuman populations and loci where reference standards are not available. Consequently, in most cases alleles need to be 'binned', with fragments assigned to allele categories according to their lengths. Alternative methods of binning include fitting observed allele lengths to bins defined by the repeat unit of the locus being considered (e.g. dinucleotide repeats being matched to a two base-pair periodicity) or simply rounding fragment lengths to the nearest base.

Although standard binning methods appear to work, the goal of fully automated binning still seems some way off. For example, one study using the program GENOTYPER to place 816 alleles from 12 loci in 38 samples (Ginot *et al*. 1996) reported an error rate of ~5%, far higher than with manual approaches (Hoffman & Amos 2005a). Similarly, in a study using a commercial panel of 400 human primer pairs and 60 nonoptimized fine-mapping markers (Ewen *et al*. 2000), binning errors made by GENOTYPER accounted for 21.05% and 39.62% of call errors respectively. Elsewhere, Weeks *et al*. (2002) compared results from two laboratories and found that 82.8% of discrepancies in dinucleotide repeat microsatellite scores and 50% of tetranucleotide

Correspondence: William Amos, Fax: +44 1223 336676; E-mail: w.amos@zoo.cam.ac.uk.
Present address: Angela Frodsham, Cambridge Genetics Knowledge Park, Public Health Genetics Unit, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK

scores were due to different decisions about placing alleles in bins. Unfortunately, the overwhelming majority of studies that report and/or quantify binning errors belong to the field of human genetics, bringing two disadvantages: first, these results may well not reach the broader molecular ecology community, and second, they may underestimate the problem because human systems are by and large well-optimized.

Binning errors may arise in many ways, but the most important and often overlooked problem is that DNA fragment mobility depends both on length and on sequence (Rosenblum *et al*. 1997; Wenz *et al*. 1998). Thus, DNA fragments with a high GC content migrate differently compared with equivalent length pieces having a low GC content. Consequently, depending on the GC content of the flanking sequence, alleles may show periodicities that are close to, but not exactly the same as the underlying repeat unit. At loci where the effective repeat length differs more than negligibly from the nominal repeat length, automatic fixed binning will often place bin boundaries within rather than between the clusters of lengths that represent each allele. Typically, predetermined bin widths work nicely at one end of the length distribution but poorly at the other end (Ewen *et al*. 2000).

In addition to the problems of fixed-sized bins, there exist at least two further ways in which the binning process can be disrupted. First, mutations in the flanking sequence can create heterogeneity in the effective length of an allele, thereby broadening the length distribution and reducing the accuracy of binning. Second, while most mutations appear to involve changes of whole numbers of repeats, occasionally alleles change in length by other sizes, for example, a dinucleotide locus losing or gaining a single nucleotide. Such changes present a challenge for the binning process (Ewen *et al*. 2000) because the alleles they create may lie on what would otherwise be a valid bin boundary.

Poor binning has a number of potentially detrimental consequences. First, even small numbers of resulting errors can seriously impact individual identification (Creel *et al*. 2003; Waits & Leberg 2003) and parental exclusion (Hoffman & Amos 2005a, b). Moreover, by splitting fragments that in reality represent one allele into two different length classes, the validity of allele frequency estimates will be undermined. This will adversely affect studies of population structure and distort estimates of key parameters such as heterozygosity. In the worst-case scenario, the outcome of binning will depend on the particular gel or experiment being run. For example, length estimates can vary with the speed at which polymerase chain reaction products run through a gel or capillary relative to the standard fragments, and this can in turn vary with stochastic variation in ambient temperature and with different qualities of the separation matrix (Ghosh *et al*. 1997; Rosenblum *et al*. 1997; Bruland *et al*. 1999; Williams *et al*. 1999; Davidson & Chiba 2003). When this happens, an allele run on one day may be consistently scored as length X when exactly the same allele in the same set of samples run on a different day would be scored as $X \pm 1$ repeat or base pair. Since samples are often analysed in batches from similar times and places, this has the potential to create artefactual genetic differences.

Most of the problems described above can be (and often are) addressed by careful manual scrutiny and adjustment where poor binning is apparent. However, in our experience this is not always the case, sometimes because the operator is unaware that automated binning is far from foolproof, and sometimes because of time constraints. Even where manual checking is carried out, the process is often tedious and time-consuming. To improve both the accuracy and speed of the binning process, we have therefore developed a simple algorithm that combines a flexible approach to finding the effective repeat unit size with an output that allows rigorous checking for any other problems such as intermediate allele sizes. To test the approach, we used a published data set from a whole genome screen in humans for hepatitis B.

Materials and methods

We have designed a program, 'FLEXIBIN' to implement automated binning. The program is available on Bill Amos' website (http://www.zoo.cam.ac.uk/zoostaff/amos/) and can also be obtained by contacting w.amos@zoo.cam.ac.uk. The basic algorithm is simple and has been coded in Visual Basic as an Excel Macro. First, to ensure that allele lengths reflect as much as possible the number of repeats they contain, all lengths are first trimmed to a length $X - \min + 3$, where X is the original length, min is the minimum allele length recorded and the value of 3 is added to eliminate problems of negative allele lengths (see below). The next step is to explore all likely binning patterns. In terms of actual mobility on a gel, each dinucleotide repeat unit will contribute approximately, but seldom exactly, 2 base pairs (bp). We refer to this as the effective repeat unit length, EL, assumed to lie between 1.7 to 2.3 bp for dinucleotides. In addition to some integer number of repeats, there is also likely to be a remainder, termed the offset, O. Ideally, all alleles will have an expected mobility of $O + n\mathrm{EL}$, where n is the number of repeat units.

To find the best bin locations for a set of observed mobilities, we exhaustively explore all possible combinations of values for O and EL. For each combination, the program considers every observed fragment in turn and determines the nearest expected length, equivalent to the centre of the bin in which that allele would be placed. As a measure of fit, the value $|\mathrm{OL} - \mathrm{EL}|$ is stored, where OL is the observed length. Over all alleles, the goodness of fit to the

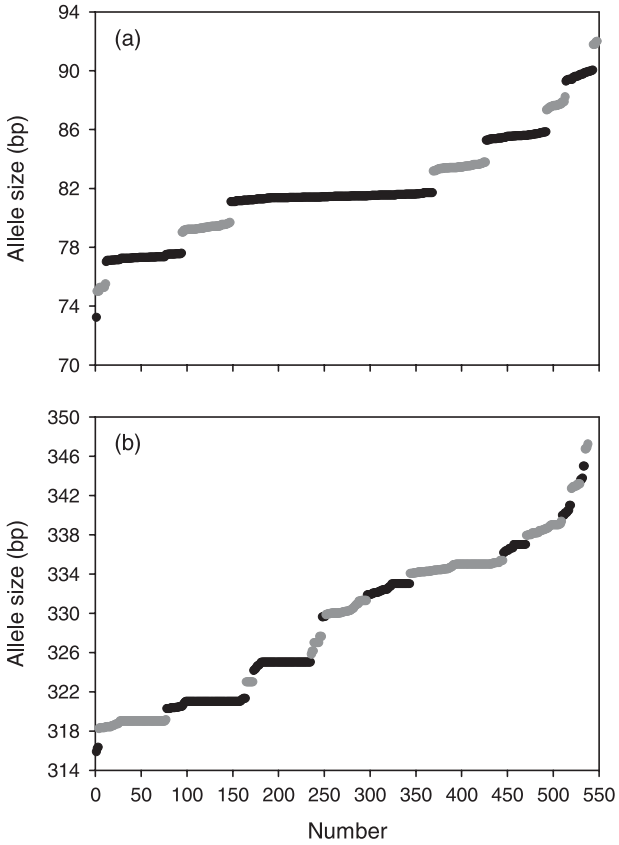

Fig. 1 Representative cumulative allele length distributions. (a) Pattern obtained from a standard locus (d7s798) showing a clear, discrete allele size distribution; (b) one of the worst cases found (locus d7s2465), showing indistinct allele size classes where even manual binning would struggle to identify allele class boundaries. Alternate allele classes as inferred are shown in contrasting colours.

current binning parameters is taken as the sum of the squared deviations between the observed length and the nearest bin centre. The smaller this sum, the greater is the tendency for fragments to lie towards the middle rather than the edge of each bin.

To find the best binning parameter, the program steps through all possible parameter combinations, seeking to find the set that achieves the best fit. For maximum resolution, the search is conducted in two phases, one coarse and then a second, fine-grained search centred on the parameter values generated by the coarse search. When the best-fit values are found, all alleles are replaced with their repeat unit equivalents and an output is produced containing summary statistics, including the standard deviation of alleles in each bin, the mean and expected lengths of alleles in each bin, the estimated size of offset and the best-fit effective repeat length. To aid interpretation, a graphical output is also generated in which alternate alleles are colour-coded and plotted as a cumulative distribution. Any misplaced alleles are then spotted with ease (Fig. 1).

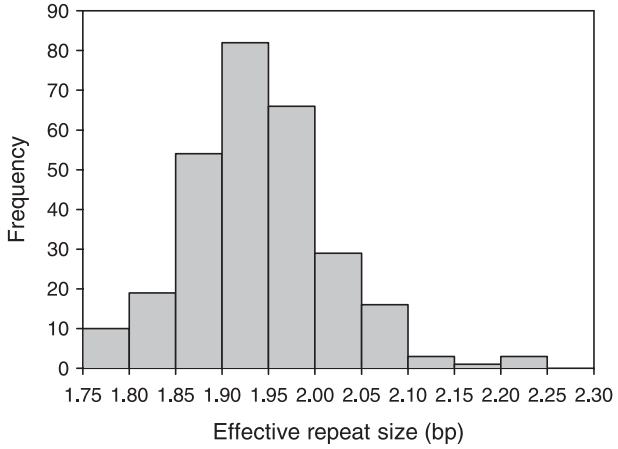

Fig. 2 Frequency distribution of effective repeat sizes obtained for 276 human autosomal dinucleotide markers.

The data set

As a test of the binning process, data from a whole genome screen for hepatitis B were analysed. These data include raw allele lengths for 318 individuals typed for 276 autosomal markers. All samples were run on an ABI 377 machine with the manufacturer's internal size standards.

Results

Overall, the binning process experienced few problems, and most alleles were binned satisfactorily. Figure 1(a) illustrates a standard, accurate binning, with a graphical output for the cumulative allele length distribution. Figure 1(b) illustrates one of the worst cases found and reveals the potential problems faced by any binning program. Although some alleles reveal to nice, discrete length distributions, others merge with each other, apparently due to allele mobility heterogeneity. There is also the problem that allele lengths do not always conform to consecutive numbers of repeat units, but instead have large length gaps that will tend to exacerbate any mismatch between the effective and applied repeat unit length.

Figure 2 depicts the observed distribution of effective repeat lengths. The range we find runs from 1.77 up to 2.23 bp, peaking clearly in the size category 1.9 to 1.95 bp, somewhat under 2 bp. Importantly, 30% of all alleles have an estimated effective repeat length of 1.9 bp or less, implying that an allele length range of 10 repeat units or more will see a proportion of alleles falling on bin boundaries if an exact 2-bp periodicity is assumed (the effective length is 0.1 bases shorter than 2 bp, so alleles 10 repeats apart will have lost one full base pair relative to a 2-bp scale).

Figure 3 gives the overall distribution of standard deviations across all allele bins at all loci. The standard

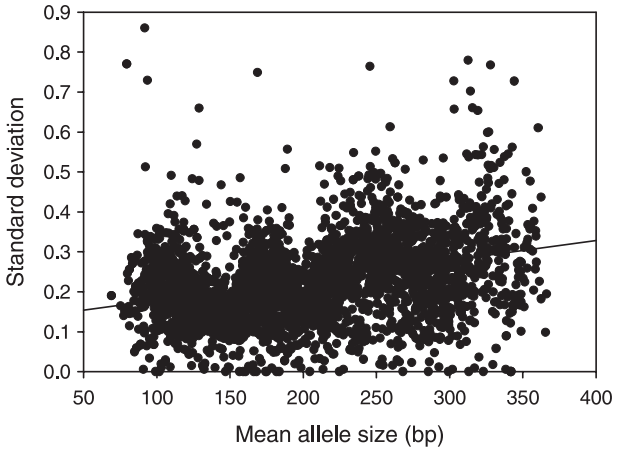

Fig. 3 Relationship between mean allele length (bp) and standard deviation summarized over 276 autosomal markers showing that the majority of alleles have a low standard deviation (< 0.4) making misclassification unlikely.

deviation for each allele bin is calculated as the square root of the variance of the mobilities of all fragments placed within that bin. Several features are apparent. First, the overwhelming majority of alleles have standard deviations less than 0.5, the level at which 95% of all alleles will be binned correctly (a standard deviation of 0.5 indicates that 95% of alleles lie within 1 bp of the bin centre). In practice, resolution will tend to be a little better than this because the deviations are taken relative to the predicted mean rather than the actual mean. Consequently, alleles whose mean does not match exactly the expected mean will tend to show a slightly inflated standard deviation. Satisfyingly, 98.3% of all alleles have a standard deviation less than 0.5, where misclassification is unlikely. The second trend is, overall, for the standard deviation to increase with increasing allele length. Finally, there is a pronounced periodicity, with standard deviations tending to be lower at lengths such as 100, 150 and 200 base pairs. These lengths correspond to the lengths of the internal standard fragments, and suggest that fragment length estimation is noticeably more accurate when the fragment lies near to a standard band. Again, this is to be expected and has been noticed previously as a tendency for between-laboratory consistency to be greater for alleles lying close to size standard bands (Jones *et al*. 1997).

Discussion

Here we describe a simple algorithm aimed at speeding up and increasing the accuracy of microsatellite allele-binning. Applied to a wide range of dinucleotide loci in humans, the results are encouraging, with a very high proportion of all alleles being apparently correctly assigned. Importantly, the generation of output statistics concerning the accuracy of the binning process allows a rapid check as to whether there are problems. We are hopeful that our approach will help to increase the level of automation in allele-calling (Bonin *et al*. 2004; Pompanon *et al*. 2005) without compromising the error rate.

The extent to which binning is viewed as a serious problem varies between studies. Where allele diversity is low and/or the loci being typed are tri- or tetranucleotides, standard approaches may be both rapid and effective. However, dinucleotide markers remain widely used for a number of reasons, including the fact that they are generally easier to clone in numbers (Zane *et al*. 2002) and were often preferred over higher order repeats historically. Even with dinucloetide repeats, experienced researchers will know to spend the necessary time making manual adjustments. Most problems therefore arise either through lack of awareness that binning is by no means trivial, or because too little time is available or spent on manual rechecking.

Perhaps the most striking aspect of this study is the range of effective repeat lengths found. The peak size estimate was in fact not two repeats, but a little under 2 bp, while the estimates ranged from as low as 1.77 bp through to an upper limit of 2.23 bp. This range corresponds approximately to the limits of the possible range enforced by the input search criteria, but this is no coincidence. During algorithm optimization, wider ranges were explored, including the use of single base periodicities. On well-behaved loci, the same answer emerged regardless of the range allowed. However, on problematic loci, reduced constraints sometimes resulted in obviously unrealistic solutions, with good fits being achieved simply by having more allele length classes. The current parameter limits appear successful for two reasons. First, the distribution of effective repeat sizes is clearly bell-shaped with well-defined tails that fit within the allowable range. Second, the fit criteria indicate that the overwhelming majority of alleles binned fall close to the bin centres and rarely approach the interbin length boundaries.

In the current study, only a tiny fraction of all alleles draw from a handful of loci revealing unusably poor binning, and most of these are easily rectifiable following manual inspection. This high success rate could be because the loci had been preselected for having relatively few problems due to, for example, alleles having both odd and even numbers of bases. However, this in no way implies that such problems do not arise or are even rare. For this reason, post hoc scrutiny remains a vital part of the whole allele calling procedure. To help alleviate this problem, possible problematic loci are flagged and a more detailed output generated.

The most widely used automated system is GENEMAPPER which uses two stages of binning. At the start of a study, allele classes are defined using a combination of an automated procedure and manual input, adjusting suggested

bin sizes and bin centres. Once all bins have been allocated, alleles are then called by finding the nearest bin centre. While we do not use GENEMAPPER ourselves, talking to colleagues reveals that three common problems seem apparent. First, none of the researchers we spoke to used the automated initial stage but instead focused on defining bins manually. Second, we also received reports of 'bin creep', a tendency for the migration rate of particular alleles to vary somewhat over time, in some cases causing a mismatch with the original bin definitions. Third, new alleles falling outside the expected range caused problems and were placed with poor accuracy. All three of these problems are largely negated by the system we use, where no reliance is placed on predefining a fixed series of bins. This offers a particular advantage for lower throughput, nonhuman systems where sample sizes are seldom large enough to be confident that all alleles have been found.

In conclusion, we demonstrate an effective algorithm for binning the output from automatic genotyping systems. Although demonstrated for dinucleotide repeats, the most challenging class, our program will also operate on data from tri- and tetranucleotides. Applied to real data, we highlight the common problem that nominal repeat length frequently fails to match the effective repeat length as estimated by an automated sequencer. We hope that our algorithm will help to decrease errors associated with automated allele-calling.

Acknowledgements

Many thanks to Josephine Pemberton and Mathieu Joron for their helpful comments about using GENEMAPPER. The genotyping work was funded by the Wellcome Trust. Joe Hoffman is currently supported by the Isaac Newton Trust and the Balfour Fund.

References

Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P (2004) How to track and assess genotyping errors in population genetic studies. *Molecular Ecology*, **13**, 3261–3273.

Bruland O, Almqvist EW, Goldberg YP, Broman H, Hayden MR, Knappskog PM (1999) Accurate determination of the number of CAG repeats in the Huntington disease gene using a sequence-specific internal DNA standard. *Clinical Genetics*, **55**, 198–202.

Creel S, Spong G, Sands JL *et al.* (2003) Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. *Molecular Ecology*, **12**, 2003–2009.

Davidson A, Chiba S (2003) Laboratory temperature variation is a previously unrecognised source of genotyping error during capillary electrophoresis. *Molecular Ecology Notes*, **3**, 321–323.

Ewen KR, bahlo M, Treloar SA *et al.* (2000) Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics*, **67**, 727–736.

Ghosh S, Karanjawala ZE, Hauser ER *et al.* (1997) Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labelled dinucleotide markers. FUSION (Finland–US Investigation of NIDDM Genetics) Study Group. *Genome Research*, **7**, 165–178.

Ginot F, Bordelais I, Nguyen S, Gyapay G (1996) Correction of some genotyping errors in automated fluorescent microsatellite analysis by enzymatic removal of one base overhangs. *Nucleic Acids Research*, **24**, 540–541.

Hoffman JI, Amos W (2005a) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, **14**, 599–612.

Hoffman JI, Amos W (2005b) Does kin selection influence fostering behaviour in Antarctic fur seals (*Arctocephalus Gazella*)? *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **272**, 2017–2022.

Jones CJ, Edwards KJ, Castaglione S *et al.* (1997) Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Molecular Breeding*, **3**, 381–390.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, **6**, 847–859.

Rosenblum BB, Oaks F, Menchen S, Johnson B (1997) Improved single-stranded DNA sizing accuracy in capillary electrophoresis. *Nucleic Acids Research*, **25**, 3925–3929.

Schlotterer C (2004) The evolution of molecular markers-just a matter of fashion? *Nature Reviews Genetics*, **5**, 63–69.

Waits JL, Leberg PL (2003) Biases associated with population estimation using molecular tagging. *Animal Conservation*, **3**, 191–199.

Weeks DE, Conley YP, Ferrell RE, Mah TS, Gorin MB (2002) A tale of two genotypes: consistency between two high-throughput genotyping centres. *Genome Research*, **12**, 430–435.

Wenz HM, Robertson JM, Menchen S *et al.* (1998) High-precision genotyping by denaturing capillary electrophoresis. *Genome Research*, **8**, 69–80.

Williams LC, Hegde MR, Herrera G, Stapleton PM, Love DR (1999) Comparative semi-automated analysis of (CAG) repeats in Huntington disease gene: use of internal standards. *Molecular and Cellular Probes*, **13**, 283–289.

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology*, **11**, 1–16.