Journal of Heredity 2009:100(3):394–397
doi:10.1093/jhered/esn110
Advance Access publication January 6, 2009

Computer Note

MsatAllele_1.0: An R Package to Visualize the Binning of Microsatellite Alleles

Filipe Alberto

From the Centro de Ciências do Mar do Algarve, Centro de Investigação Marinha e Ambiental-Laboratório Associado, Universidade do Algarve, Campus de Gambelas, 8005-139 Faro, Portugal (Alberto).

Address correspondence to F. Alberto at the address above, or e-mail: falberto@ualg.pt.

MsatAllele is a computer package built on R to visualize and bin the raw microsatellite allele size distributions. The method is based on the creation of an R database with exported files from the open-source electropherogram peak-reading program STRAND. Contrary to other binning programs, in this program, the bin limits are not fixed and are automatically defined by the data stored in the database. Data manipulation and graphical functions allow to 1) visualize raw allele size variation, allowing the detection of potential scoring errors, strange bin distributions, and unexpected spacing between the bins; 2) bin raw fragment sizes and write bin summary statistics for each locus; and 3) export genotype files with the resulting binned data.

Key words: *bin, database, microsatellite markers, plotting, scoring errors*

Genomic microsatellites, also called simple sequence repeats or short tandem repeats, are powerful tools commonly used to characterize the neutral (and also more recently selected; e.g., Vasemagi et al. 2005) genetic diversity. Over the last 15 years, microsatellites have been used as the tool of choice to address many population genetics and demographic questions (Estoup and Angers 1998), test ecological and evolutionary hypotheses (Estoup and Angers 1998), and quantify the genetic diversity components of relevant interest for conservation biology (Allendorf and Luikart 2007). The high sensitivity of the polymerase chain reaction-based microsatellite analysis was not only of great benefit in forensics but also opened completely new research areas such as the analysis of samples with limited DNA amounts or degraded DNA (e.g., feces, museum material, and mark recapture methods). However, microsatellite analyses can be affected by scoring errors, a fact that is not sufficiently emphasized (Paetkau 2003; Pompanon et al. 2005; Dewoody et al. 2006; Amos et al. 2007). Scoring microsatellite alleles can be complicated when there is a high number of alleles per locus, which is often the case observed. Additionally, the stutter bands and changes of the regular pattern of mutation (e.g., 1-bp apart alleles instead of a regular addition or subtraction of one or more microsatellite repeat motif; Ewen et al. 2000) can result in the accumulation of an important number of scoring errors (see Dewoody et al. 2006). Apart from the technical problems associated with the microsatellite scoring errors, the human factors, such as inexperienced reader, ambiguity, and subjectivity are largely ignored in the literature. One of the causes of error is erroneous binning (the process of assigning an integer allele code to the continuous fragment size value obtained from the sequencer) owing to ambiguous peak calling, different readers, or different amplification or electrophoresis conditions (Davison and Chiba 2003).

Most allele binning softwares, although offering a supposedly fully automated analyses, do not provide a way of visualizing the global distribution of fragment sizes for a given locus, with the exceptions of the downstream methods provided by ALLELOGRAM (Manaster 2002) and FLEXIBIN (Amos et al. 2007). Yet, a correct binning method still relies largely on a careful and consistent manual scoring of electropherograms, and hence, a downstream visualization of the global distributions of raw fragment sizes is a very convenient method to identify calling problems, such as stutter or ambiguous peak calling.

Several packages offer binning methods; some like GENEMAPPER can be fully automated if a system is properly optimized and calibrated, which unfortunately is not always the case for most species. Bins' limits are set arbitrarily by the user in GENEMAPPER, a process that can be cumbersome in cases with high number of alleles per locus and many loci or when there are alleles differing by 1 bp (Ghosh et al. 1997; Ewen et al. 2000). Most important, these and other common peak calling and allele binning programs have expensive licenses and heavy hardware requirements to be easily installed on any computer, often restricting the number of computers available that can be used to read genotyping data.

We developed an R package (R Development Core team 2007), MsatAllele, to visualize and bin microsatellite fragment sizes by performing several data manipulations on an R database built with the raw fragment size data exported from STRAND (http://www.vgl.ucdavis.edu/informatics/strand.php). The combination of these 2 open-source programs allows users to read and call microsatellite allele data, free of cost and on any ordinary computer.

R is an open-source software for the statistical analysis and complex computations with numerous graphical applications and programming language (R Development Core team 2007). To use this package, it is necessary to know little about R and follow the detailed instructions in

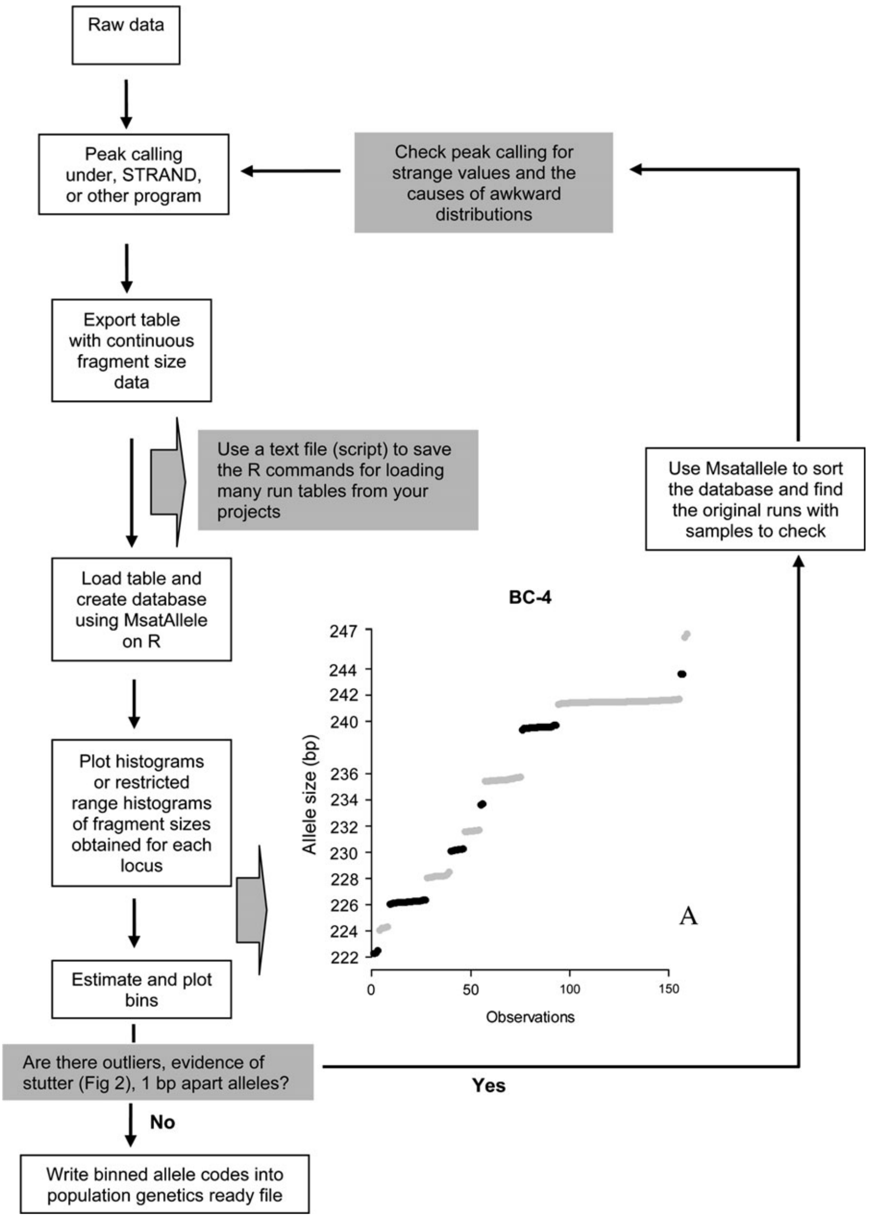
Raw data
Peak calling under, STRAND, or other program
Check peak calling for strange values and the causes of awkward distributions
Export table with continuous fragment size data
Use a text file (script) to save the R commands for loading many run tables from your projects
Use Msatallele to sort the database and find the original runs with samples to check
Load table and create database using MsatAllele on R
Plot histograms or restricted range histograms of fragment sizes obtained for each locus
Estimate and plot bins
Are there outliers, evidence of stutter (Fig 2), 1 bp apart alleles?
Yes
No
Write binned allele codes into population genetics ready file
BC-4
Allele size (bp)
247
244
242
240
236
234
232
230
228
226
224
222
0
50
100
150
Observations
A

Figure 1. Data analysis flow from the sequencer raw data output to peak calling, exporting fragment size data (continuous unbinned data), loading data into R, and constructing a database. After an iterative visualization of the distribution of the allele sizes called and rechecking the suspicious observations back in the peak-scoring software, a final file with the bin alleles (integer data) can be exported from R.

the MsatAllele manual. The MsatAllele package contains functions to 1) load exported table files containing continuous fragment size data and build an R database that can be added to larger, previously recorded databases; 2) plot interactive cumulative distributions of fragment size observations providing a global visualization of size differences among bin distributions and a quick method to locate strange observations; 3) plot full and restricted range histogram distributions obtained for each locus; 4) provide a bin algorithm fully dependent on the distribution properties of the observed fragment sizes, that is, does not rely on the user-defined bin limits; 5) write summary statistics for each bin of a given locus; 6) sort the database for a given locus and range, which allows easy traceback of particular outlier samples to its original electropherogram file; and 7) write files with binned data ready for population genetics.

Binning Algorithm

MsatAllele provides a binning method that uses all the raw fragment size information for a given locus stored on an

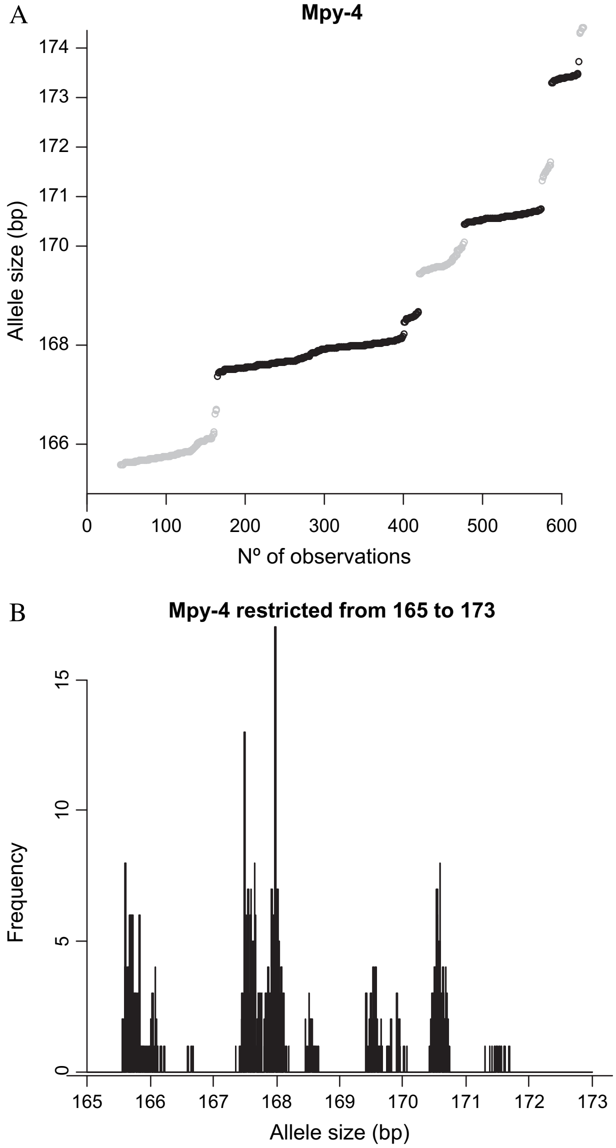

Figure 2. Two different visualizations of raw fragment size distributions that can be produced by MsatAllele: (**A**) cumulative fragment size distribution, the integer values in the *y* axis are the bin codes, the points color alternates from gray to black whenever the bin attributed changes; (**B**) restricted histogram, useful to visualize extremely variable locus. When dealing with problematic loci, histogram visualization is more efficient to detect problems, such as the bimodal distributions inside bins, shown in this example. The causes can be stutter or variable electropherogram pattern, causing ambiguous calling of these alleles.

R database. Observations are binned into allele codes by searching the database and detecting their respective bin distribution. In brief, the process of binning each observation starts by extracting a vector from the database containing all the observations within ±0.8 bp of that raw fragment size. The value of 0.8 bp is assumed to slightly overestimate the expected standard deviation inside a bin. Because a given observation can fall at the tail of the distribution, 0.8 bp guarantees that all the observations for that bin are extracted, although it is possible that the observations from more than one bin distribution fall inside this vector, in cases where 1-bp separated alleles are present. Thus, to find the distribution to which the fragment belongs, the method checks for breaks in the vector, that is, the consecutive observations separated by >0.4 bp. If more than one distribution is present, then the one with more observations is retained. Furthermore, to obtain the bin code, the median of that distribution is rounded to an integer value.

Brief Illustration

MsatAllele is a downstream binning method after peak calling is performed, and the analysis flow is depicted in Figure 1. The function read.ah.file reads STRAND exported files with fragment size data into MsatAllele. Other output formats can be easily converted to a simple input text file readable by read.frag.sizes function. Once the fragment sizes are imported into R, they can be merged into the existing databases using simple R functions (e.g., rbind). It is convenient to maintain a script file recording the commands loaded into R with the names of the input files containing the raw fragment size data. When the database is complete, the AlleleHist or restrict.hist functions can be used to plot global and restricted histograms of the fragment size distributions per locus (Figure 1). Cumulative fragment size distribution plots are also available with AlleleCum. The resulting plots can be used to check the plots for outliers, rare alleles, bins with high standard deviation, unexpected spacing between consecutive bin distributions, and bimodal distributions (potentially caused by the heterogeneous calling of alleles, Figure 2). Suspicious observations and potential scoring errors can be identified and traced back to their original electropherograms by printing a sublist of the database sorted by size, using the subdataBase function, or interactively by clicking on the graph with function getpoints. MsatAllele visualization can also assist readers (particularly, the first-time readers of a previously characterized loci set) during the scoring stage of new samples for loci recorded in the existing databases. When checks are completed, a file with the binning results, ready for population genetics statistical analyses, can be written using write.PG.file.all or write.PG.file.loc functions, for 2 and 1 column per locus formats, respectively. The package MsatAllele and its detailed manual can be downloaded from http://www.ccmar.ualg.pt/maree/software.php.

Funding

Fundação para a Ciência e Tecnologia pos-doctoral grant (SFRH/BPD/14945/2004 to F.A.); MEGIKELP (PTDC/MAR/65461/2006).

Acknowledgments

The author thanks Ester Serrão for her comments on this work and Nelson Coelho for generating the extremely variable data that caused the need for this package. He also expresses his gratitude to Sónia Massa for the STRAND lessons.

References

Allendorf FW, Luikart G. 2007. Conservation and the genetics of populations. Oxford: Blackwell Publishing.

Amos W, Hoffman JI, Frodsham A, Zhang L, Best S, Hill AVS. 2007. Automated binning of microsatellite alleles: problems and solutions. Mol Ecol Notes. 7:10–14.

Davison A, Chiba S. 2003. Laboratory temperature variation is a previously unrecognized source of genotyping error during capillary electrophoresis. Mol Ecol Notes. 3:321–323.

Dewoody J, Nason JD, Hipkins VD. 2006. Mitigating scoring errors in microsatellite data from wild populations. Mol Ecol Notes. 6:951–957.

Estoup A, Angers B. 1998. Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. In: Carvalho GR, editor. Advances in molecular ecology. Amsterdam (the Netherlands): IOS Press. p. 55–86.

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. Am J Hum Genet. 67:727–736.

Ghosh S, Karanjawala ZE, Hauser ER, Ally D, Knapp JI, Rayman JB, Musick A, Tannenbaum J, Te C, Shapiro S, et al. 1997. Fusion Study Group. 1997. Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. Genome Res. 7:165–178.

Manaster CJ. 2002. Allelogram: a program for normalizing and binning microsatellite genotypes [Internet]. Available from: http://code.google.com/p/allelogram/.

Paetkau D. 2003. An empirical exploration of data quality in DNA-based population inventories. Mol Ecol. 12:1357–1387.

Pompanon F, Bonin A, Bellemain E, Taberlet P. 2005. Genotyping errors: causes, consequences and solutions. Nat Rev Genet. 6:847–859.

R Development Core Team. 2007. R: a language and environment for statistical computing [Internet]. Vienna (Austria): R Foundation for Statistical Computing. ISBN 3-900051-07-0. Avalilable from: URL http://www.R-project.org.

Vasemagi A, Nilsson J, Primmer CR. 2005. Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (Salmo salar L.). Mol Biol Evol. 22:1067–1076.

Received November 4, 2008; Accepted December 1, 2008

Corresponding Editor: C. Scott Baker