Journal of Heredity
doi:10.1093/jhered/esq111

Computer Note

MICROSATELIGHT—Pipeline to Expedite Microsatellite Analysis

FERRAN PALERO, FERNANDO GONZÁLEZ-CANDELAS, AND MARTA PASCUAL

From the Evolutionary Genetics, Institute of Science and Technology Austria (ISTA), Am Campus 1, A–3400 Klosterneuburg, Austria (Palero); the Unidad mixta "Genómica y Salud" Centro Superior de Investigación en Salud Pública-Universidad de Valencia/Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Valencia, Spain (Fernando); the CIBER de Epidemiología y Salud Pública, Valencia, Spain (Fernando); and the Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain (Marta).

Address correspondence to F. Palero at the address above, or e-mail: fpalero@ist.ac.at.

MICROSATELIGHT is a Perl/Tk pipeline with a graphical user interface that facilitates several tasks when scoring microsatellites. It implements new subroutines in R and PERL and takes advantage of features provided by previously developed freeware. MICROSATELIGHT takes raw genotype data and automates the peak identification through PeakScanner. The PeakSelect subroutine assigns peaks to different microsatellite markers according to their multiplex group, fluorochrome type, and size range. After peak selection, binning of alleles can be carried out 1) automatically through AlleloBin or 2) by manual bin definition through Binator. In both cases, several features for quality checking and further binning improvement are provided. The genotype table can then be converted into input files for several population genetics programs through CREATE. Finally, Hardy–Weinberg equilibrium tests and confidence intervals for null allele frequency can be obtained through GENEPOP. MICROSATELIGHT is the only freely available public-domain software that facilitates full multiplex microsatellite scoring, from electropherogram files to user-defined text files to be used with population genetics software. MICROSATELIGHT has been created for the Windows XP operating system and has been successfully tested under Windows 7. It is available at http://sourceforge.net/projects/microsatelight/.

Key words: *bioinformatics, population genetics, SSR, software,*

Manual scoring of amplified fragment length polymorphisms (AFLP) and microsatellite markers is prone to data entry errors, time intensive, and subjective (McGreevy et al. 2009). This time-consuming task directly affects the quality of the final data set, particularly when sampling includes many markers and a large number of individuals. Although software for automated scoring of allele size is available (AFLPScore v1.3: Whitlock et al. 2008; TANDEM v1.07: Matschiner and Salzburger 2009; RawGeno: Arrigo et al. 2009), no freeware allows for simultaneous analysis of multiple marker groups. Most freely available programs cover just a single step in the process (MsatAllele v1.01: Alberto 2009), and in some cases, the portability is reduced by working on a very specific program environment (FLEXIBIN: Amos et al. 2007) or demanding a particular input file (Allelogram: Morin et al. 2009).

MICROSATELIGHT is a Perl/Tk graphical user interface that facilitates several tasks for microsatellite scoring. It does so by implementing new subroutines in PERL and R and using features provided by previously developed freeware (R Development Core Team 2009). MICROSATELIGHT provides new capabilities to reduce the time spent by the user on routine tasks (e.g., repeating analyses for different multiplex groups) and by automating file conversion and software calling steps. MICROSATELIGHT has been created for the Windows XP operating system and has been successfully tested under Windows 7.

A single tab-delimited text file is needed to provide information on the multiplex group to which each marker belongs, locus name, minimum and maximum sizes attained, fluorochrome used for labeling the marker (e.g., G = HEX; B = FAM; Y = NED), and repeat pattern for each locus (see Supplementary Material). A sample file is included with the MICROSATELIGHT distribution (param.txt).

Running MICROSATELIGHT is a 7-step process (Figure 1):

1. Automatically running PeakScanner v1.0 for extracting peak size data from .fsa files (Sizing_Table.txt);
2. Trimming the Sizing_Table.txt file and assigning peaks to different microsatellite markers according to their multiplex group, fluorochrome type, and size range by using the PeakSelect subroutine (Selected.txt);
3. Creating a 2-column genotype table with observed allele sizes through the GenotypeTable subroutine (Observed_genotypes.txt).

Once the Observed_genotypes.txt file is obtained, it is possible to carry out the allele calling step in MICROSATELIGHT.

4. Automatically, through the least-square minimization algorithm of Idury and Cardon (1997) as implemented in AlleloBin (Prasanth et al. 2006);
5. Manually, defining bin limits through Binator.

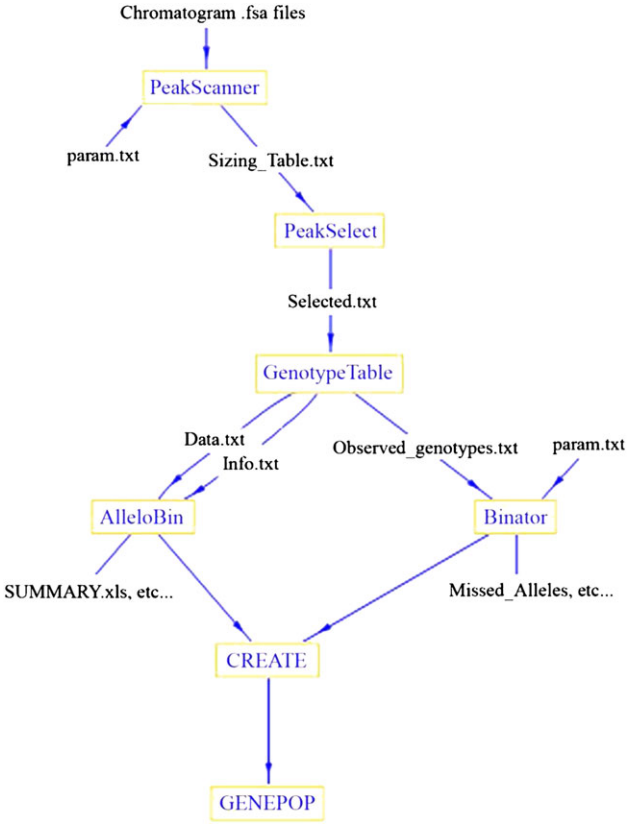

Figure 1. MICROSATELIGHT pipeline flowchart.

In both cases, several features for quality checking and further binning improvement are provided. AlleloBin output includes parameter estimates for the least-squares procedure and provides a measure of fractional repeat patterns called "Allelic Drift." Binator allows for identification of missed alleles by comparing selected peaks with raw data. Furthermore, Binator output includes a list of alleles that do not follow the known repeat length pattern and a list of problematic chromatogram files.

6. After obtaining the Binned_Genotypes file, the user may run CREATE (Coombs et al. 2008) to convert the genotype table into input files for several population genetics programs.
7. Finally, GENEPOP can be called within MICROSATELIGHT in order to carry out further analyses (e.g., test for departure from Hardy–Weinberg equilibrium).

This process provides a balance between the efficiency and consistency of automated allele-calling software and the accuracy provided by human inspection in detecting novel alleles outside the expected size range of a locus. MICROSATELIGHT helps the user to identify potential mistypes due to stutter patterns or large-allele dropout by allowing comparison of raw data and selected peaks, collecting summary statistics on bin definition and later testing for Hardy–Weinberg equilibrium with GENEPOP. It is important to note that both the Binator and the AlleloBin methods allow interbin distances to vary across alleles. Allowing interbin variability yields accurate binning even under difficult conditions such as bimodal distributions associated with plus-A amplification.

MICROSATELIGHT graphical user interface allows the user to define multiple marker groups or sample names through several dialog boxes. Moreover, it automatically creates input files for specific software, facilitating integration of different tasks within a single framework. Therefore, MICROSATELIGHT provides an easy-to-use environment for dealing with allele-sizing data, and it will help to decrease errors associated with genotyping. MICROSATELIGHT installer and precompiled binaries are available at: http://sourceforge.net/projects/microsatelight/.

Supplementary Material

Supplementary material can be found at http://www.jhered.oxfordjournals.org/.

Funding

Ministerio de Educación y Ciencia (CGL2006-13423, CTM2007-66635). M.P. and FP are part of the research group 2009SGR-636 of the Generalitat de Catalunya. F.P. acknowledges an EU-Synthesys grant (GB-TAF-4474).

Acknowledgments

Thanks to José Gabriel Segarra-Moragues (Centro de Investigaciones sobre Desertificación) for sending us pictures with several types of stuttering and Pedro Simões and Gemma Calàbria (Universitat de Barcelona) for testing this software. Finally, thanks are due to 2 anonymous referees for their valuable comments. These comments certainly helped to improve the manuscript.

References

Alberto F. 2009. MsatAllele_1.0: an R package to visualize the binning of microsatellite alleles. J Hered. 100:394–397.

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.

Arrigo N, Tuszynski JW, Ehrich D, Gerdes T, Alvarez N. 2009. Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring. BMC Bioinformatics. 10:33.

Coombs JA, Letcher BH, Nislow KH. 2008. CREATE: a software to create input files from diploid genotypic data for 52 genetic software programs. Mol Ecol Res. 8:578–580.

Idury RM, Cardon LR. 1997. A simple method for automated allele binning in microsatellite markers. Genet Res. 7:1104–1109.

Matschiner M, Salzburger W. 2009. TANDEM: integrating automated allele binning into genetics and genomics workflows. Bioinformatics. 25:1982–1983.

McGreevy Jr., TJ, Markert J, Grear JS, Nacci DE. 2009. Bridging the gap between large-scale data sets and analyses: semi-automated methods to

facilitate length polymorphism scoring and data analyses. Presented at American Genetic Association Annual Symposium: The Genetics and Genomics of Environmental Change Jun 08–11; Providence, RI.

Morin PA, Manaster C, Mesnick SL, Holland R. 2009. Normalization and binning of historical and multi-source microsatellite data: overcoming the problems of allele size shift with ALLELOGRAM. Mol Ecol Res. 9:1451–1455.

Prasanth VP, Chandra S, Hoisington DA, Jayashree B. 2006. AlleloBin—a program for allele binning of microsatellite markers based on the alogirithm of Idury and Cardon, 1997. Delhi (India): ICRISAT. International Crops Research Institute for the Semi-Arid Tropics.

R Development Core Team 2009. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. [cited 2010 Oct 15]. Available from: http://www.R-project.org

Whitlock R, Hipperson H, Mannarelli M, Butlin RK, Burke T. 2008. An objective, rapid and reproducible method for scoring AFLP peak-height data that minimizes genotyping error. Mol Ecol Res. 8:725–735.

Received March 6, 2010; Revised September 21, 2010; Accepted October 5, 2010

Corresponding Editor: Scott Baker