Benchmarks

Standardized SSR allele naming and binning among projects

Dennis L. Deemer and C. Dana Nelson

U.S. Forest Service, Southern Research Station, Southern Institute of Forest Genetics, Saucier, MS, USA

BioTechniques 49:835-836 (November 2010) doi 10.2144/000113540

Keywords: short simple repeats; microsatellites; alleles; bins; standardization

Supplementary material for this article is available at www.BioTechniques.com/article/113540.

Simple sequence repeats (SSRs) have proven to be extremely valuable DNA markers for genetic mapping and population genetic analyses. However, data collected across laboratories or even within laboratories are difficult to combine due to challenges in standardizing allele names, especially for nonmodel systems. Here we provide a new approach for standardizing SSR allele names that combines several previously recognized components for standardization, including reference samples/alleles, cumulative binsets, static between-allele spacing, and interval allele naming.

Due to a variety of factors affecting denaturing capillary electrophoresis, such as polymer lot, buffer concentration, array quality, ambient laboratory temperature, and fluorescent label, variation in PCR product sizing with internal size standards does occur (1–3). Because of this variation, simple sequence repeat (SSR) marker alleles sized, binned, and named “on the fly” are subject to inconsistency in allele naming, detracting from the inherent value of fragment length–based, sequence-specific markers (4,5). Here we describe an approach for PCR fragment binning and naming that allows for consistent, standardized allele identification between runs and projects. The approach incorporates a combination of reference samples providing reference alleles, cumulative binsets (i.e., list of all observed alleles, providing allele or bin name, expected bin center, and width) with static bin spacing and interval bin naming, and prediction equations for adjusting bin centers when changing dye labels. The latter feature is important for efficient PCR multiplexing and product pooling for electrophoresis while facilitating standardized allele naming among projects. In addition, by not forcing SSR alleles into motif repeat-length classes, our binning approach allows for alleles that

differ by non-repeat length distances (6,7). While similar in several respects to recently described methods (4,5,8), our approach offers advantages for working with larger numbers of markers (e.g., the hundreds of markers used in genetic mapping applications), independent discovery and naming of new alleles, and flexibility in dye labeling between experiments.

The basic protocol, as implemented with ABI 3130xl and 3730xl instruments (Applied Biosystems, Foster City, CA, USA) and GeneMapper software (Applied Biosystems), is outlined in Figure 1. For a run of 96 samples (i.e., six capillary fills on a 3130xl or one fill on a 3730xl), four reference samples are loaded along with the test samples. The reference samples consist of different, unrelated template DNAs amplified with the same SSR primer pairs and labeled the same as used for the test samples. Reference DNA templates, providing known SSR marker genotypes (i.e., reference alleles), are maintained in the laboratory and shared indefinitely if possible. Although a common set of reference templates is preferred, any template scored with this approach can serve as a reference. Prior to the run, panels and binsets files are formed (using script 1; see Supplementary Materials),

describing which markers are used in each panel (i.e., markers pooled together prior to electrophoresis) and the known bins (i.e., alleles) for each marker. Following capillary electrophoresis, the panels and binsets files, along with the electrophoresis output files, are imported to GeneMapper (session 1). In GeneMapper, the panels and binsets are applied; then, reference samples’ alleles are inspected and called by their previously assigned allele (i.e., bin) name and then exported in a genotypes table for script 2 (see Supplementary Materials). Script 2 models the observed reference allele sizes against the expected sizes (provided by original binsets file) and then applies the derived prediction equations to adjust the bin centers (forms the adjusted binsets file). The adjusted binsets file is imported to GeneMapper (session 2) and subsequently used to automatically call all alleles for the project. The analyst inspects these calls, paying special attention to allele calls that fall outside of the adjusted bins. These calls are usually new alleles that the analyst names and calls using the rules discussed herein. Final allele calls are exported along with the final binsets. The final allele call file can be used by various scripts to format the genotype data for specific genetic analysis programs. The final binsets file is appended to the laboratory’s cumulative binsets file for use by script 1 in forming the initial binsets file for the next project.

Five key features work together to make this approach novel and successful. First, the use of reference samples within runs provides fixed points to calibrate the test samples’ alleles to the known reference alleles. Second, cumulative binsets provide a convenient means to keep up with the current allele population for each marker. Third, static between allele spacing works well because base pair differences between alleles are more reliable than overall allele size estimates (see Figure 1 of Reference 5). Fourth, interval allele naming allows some information to be contained and referred to within the allele name—an approximate size of the fragment, but more importantly, its relationship to other alleles in terms of estimated size differences. For example, if allele 137 is estimated at 134.7 bp in a run, then for that run we would expect allele 139 to be at 136.7 ± 0.2 bp (or 2 bp larger than allele 137) and allele 143 to be at 139.7 ± 0.2 bp (or 6 and 4 bp larger than alleles 137 and 139, respectively). This allows for new alleles to be independently identified and compatibly named. Fifth, dye-dye migration prediction equations (empiri-

Figure 1. Data flow for standardized allele naming and binning protocol. GeneMapper, session 1: (1) import files, (2) call reference samples' alleles using previously identified/assigned allele (i.e., bin) names, (3) export reference samples' data as genotypes table for script 2. GeneMapper, session 2: (1) import adjusted binsets file, (2) call test samples' alleles using adjusted binsets, (3) export final binsets file for script 1 and final allele calls as genotypes table for data analysis.

cally determined in-house) improve the initial binsets when using different dyes between runs (or between projects) with the same marker. This along with the m13 labeling system (9) allows for complete flexibility in developing PCR product pools for maximizing throughput in capillary electrophoresis.

Standardized SSR fragment binning and allele naming is required for developing robust marker panels for various applications in DNA fingerprinting and genetic mapping (4,5,8). The described approach offers a simple, efficient, and effective system for achieving this goal. We have used it successfully for *Pinus* and *Castanea* SSR marker sets across projects and ABI platforms and suggest that it can be easily adopted by cooperating labs to ensure compatible allele calls among population samples and mapping pedigrees. For example, in our projects where ABI platforms are not changed, ~95% of the markers are amenable to this approach with 0% miscall rates (cases where an observed allele does not fall within its adjusted bin). Minor adjustments are needed and can be easily made for nonamenable markers, such as widening a specific bin for a known variably migrating allele or using additional reference samples for a marker with a large allele size range. Changing platforms is more challenging, as we have found consistent

resolution differences between platforms. However, in one extreme case in which we changed platforms and species (not recommended), 70% of the markers yielded 0% miscalls. As expected, those markers with relatively larger allele size ranges (>50 bp) or numbers of alleles (>10) tended to suffer miscalls, apparently due both to the changes in platform resolution and species analyzed. Finally, simple computer scripts help to automate the flow of information for utilizing and updating binsets and providing these data to capillary electrophoresis machines and semi-automated allele calling software. More sophisticated binning algorithms (10,11) could be integrated with our approach (in script 2), but we find that fixed bin widths (0.4 bp) work well when all bin centers are properly adjusted as described above.

Acknowledgments

We thank our colleagues Craig Echt, Tom Kubisiak, Chuck Burdine, and Sedley Josserand (Southern Institute of Forest Genetics) for many productive discussions on how to consistently score SSR alleles. We also thank Tom, Chuck, and the Fagaceae Genomic Tools Project (National Science Foundation) for *Castanea* SSR reagents and data. We also thank Jeanne Romero-Severson and Rita Costa for providing valuable comments during presubmission review.

Competing interests

The authors declare no competing interests.

References

- Idury, R.M. and L.R. Cardon. 1997. A simple method for automated allele binning in microsatellite markers. *Genome Res.* 7:1104-1109.
- Delmotte, F., N. Leterme, and J.-C. Simon. 2001. Microsatellite allele sizing: difference between automated capillary electrophoresis and manual technique. *BioTechniques* 31:810-818.
- Hahn, M., J. Wilhelm, and A. Pingoudet. 2001. Influence of fluorophore dye labels on the migration behavior of polymerase chain reaction—amplified short tandem repeats during denaturing capillary electrophoresis. *Electrophoresis* 22:2691-2700.
- Cryer, N.C., M.G.E. Fenn, C.J. Turnbull, and M.J. Wilkinson. 2006. Allele size standards and reference genotypes to unify international cocoa (*Theobroma cacao* L.) microsatellite data. *Genet. Resour. Crop Evol.* 53:1643-1652.
- Seeb, L.W., A. Antonovich, M.A. Banks, T.D. Beacham, M.R. Bellinger, S.M. Blankenshii, M.R. Campbell, N.A. Decovich, et al. 2007. Development of a standardized DNA database for Chinook salmon. *Fisheries* 32:540-552.
- Peakall, R., S. Gilmore, W. Keys, M. Morgante, and A. Rafalski. 1998. Cross-species amplification of soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. *Mol. Biol. Evol.* 15:1275-1287.
- Glaubitz, J.C., L.C. Emebiri, and G.F. Moran. 2001. Dinucleotide microsatellites from *Eucalyptus sieberi*: inheritance, diversity, and improved scoring of single-base differences. *Genome* 44:1041-1045.
- Stephenson, J.J., M.R. Campbell, J.E. Hess, C. Kozfkay, A.P. Matala, M.V. McPhee, P. Moran, S.R. Narum, et al. 2009. A centralized model for creating shared, standardized, microsatellite data that simplifies inter-laboratory collaboration. *Conserv. Genet.* 10:1145-1149.
- Schuelke, M. 2000. An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* 18:233-234.
- Amos, W., J.I. Hoffman, A. Frodsham, L. Zhang, S. Best, and A.V.S. Hill. 2007. Automated binning of microsatellite alleles: problems and solutions. *Mol. Ecol. Notes* 7:10-14.
- Matschiner, M. and W. Salzburger. 2009. TANDEM: integrating automated allele binning into genetics and genomics workflows. *Bioinformatics* 25:1982-1983.

Received 28 June 2010; accepted 29 September 2010.

Address correspondence to C. Dana Nelson, US Forest Service, Southern Research Station, Southern Institute of Forest Genetics, Saucier, MS, 39574, USA. e-mail: dananelson@fs.fed.us

To order reprints of this article, contact: carmelitag@fosterprinting.com