Molecular Ecology Notes (2003) doi: 10.1046/j.1471-8286 .2003.00351.x

PROGRAM NOTE

MICROSATELLITE ANALYSER (MSA): a platform independent analysis tool for large microsatellite data sets

DANIEL DIERINGER and CHRISTIAN SCHLÖTTERER
Institut für Tierzucht und Genetik; Veterinärmedizinische Universität Wien, Josef Baumann Gasse 1, A-1210 Wien, Austria

Abstract

In molecular ecology the analysis of large microsatellite data sets is becoming increasingly popular. Here we introduce a new software tool, which is specifically designed to facilitate the analysis of large microsatellite data sets. All common microsatellite summary statistics and distances can be calculated. Furthermore, the MICROSATELLITE ANALYSER (MSA) software offers an improved method to deal with inbred samples (such as Drosophila isofemale lines). Executables are available for Windows and Macintosh computers.

Keywords: F_{ST}, genetic distances, microsatellites, summary statistics

Received 22 July 2002; revision received 3 October 2002; accepted 3 October 2002

Within the past decade microsatellites have developed into one of the most popular genetic markers. Large genome scans using microsatellites were initially limited to mapping studies. More recently, with the advent of completely sequenced genomes and high throughput capillary sequencers, large scale microsatellite typing is emerging as an important tool for dissection of evolutionary forces, such as selection and hybridization in natural populations (Pritchard *et al*. 2000; Kauer *et al*. 2002; Payseur *et al*. 2002; Schlötterer 2002). While laboratory methods are fully developed to keep up with these requirements, the development of analytical tools is still lagging behind. Most of the population genetic software widely used to calculate basic statistics, such as heterozygosity or F_{ST} values was originally designed for a moderate number of loci only. Hence, these programs often use their own specific input format, which sometimes requires major reformatting efforts. Furthermore, the number of loci and individuals is often limited.

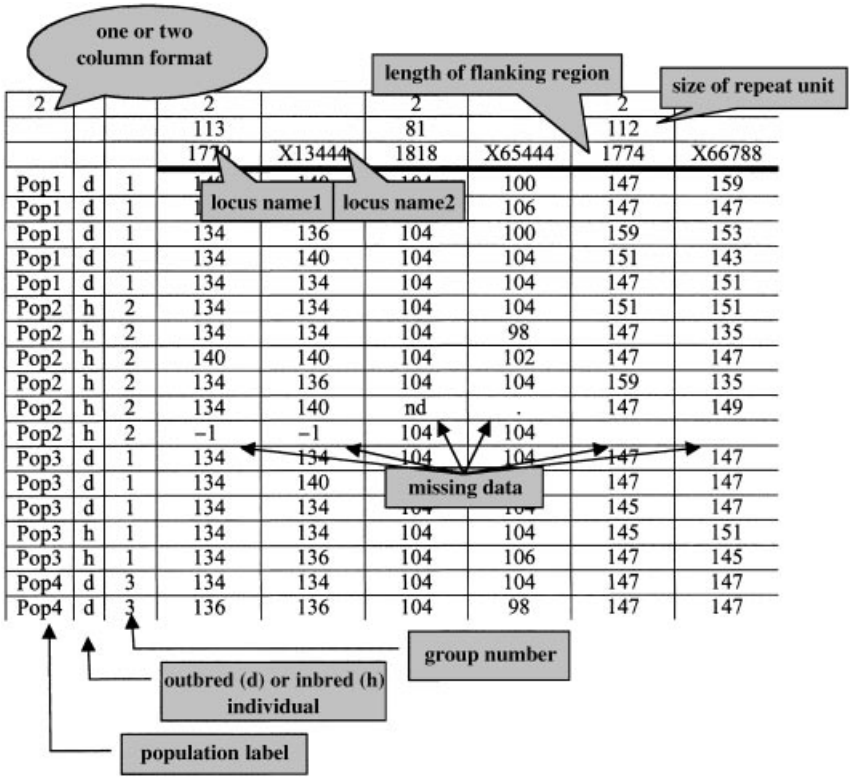
We developed the MICROSATELLITE ANALYSER (MSA) software tool specifically to overcome these limitations. In particular the MSA meets the following criteria: (a) simple input format; (b) capacity to analyse large data sets; (c) presentation of summary statistics in an open format permitting further downstream creative data mining; (d) ways of accounting for special problems encountered by the use of inbred lines (e.g. *Drosophila melanogaster*).

Correspondence: Christian Schlötterer. Fax: +43-1-25077/5693; E-mail: Christain.Schloetterer@vu-wien.ac.at

Input format. Microsatellite data are commonly stored in spreadsheet files, in which one or two columns represent each locus. Only a small number of modifications are required to transform such a file into a MSA input file. The only additional information required is an assignment of individuals to populations and the names of the loci. Optional information includes higher level grouping of populations, size of the repeat unit and number of bp flanking the microsatellite (Fig. 1).

Standard data analysis. Observed and expected heterozygosity, variance in polymerase chain reaction (PCR) product size, variance in repeat number, allele frequency distribution, number of alleles, and allele size range are the summary statistics provided by MSA. Furthermore, the expected number of alleles under the infinite allele model as well as under the stepwise mutation model are calculated (Ewens 1972; Kimura & Ohta 1975). Nine different genetic distance measurements are available either between populations or individuals. Statistical support is provided by nonparametric bootstrapping. F_{ST}, F_{IS} and F_{IT} measurements are calculated by the Weir and Cockerham method (Weir & Cockerham 1984). Significances are determined by permuting genotypes (or alleles) among groups.

Special statistics for inbred lines. Drosophila population genetics often relies on the use of isofemale lines. To account for the random loss of alleles due to inbreeding one allele is discarded (Schlötterer *et al*. 1997; Irvin *et al*. 1998; Agis & Schlötterer 2001; Kauer *et al*. 2002). Unless sample sizes are



2			2		2		2	
			113		81		112	
			1770	X13444	1818	X65444	1774	X66788
Pop1	d	1				100	147	159
Pop1	d	1				106	147	147
Pop1	d	1	134	136	104	100	159	153
Pop1	d	1	134	140	104	104	151	143
Pop1	d	1	134	134	104	104	147	151
Pop2	h	2	134	134	104	104	151	151
Pop2	h	2	134	134	104	98	147	135
Pop2	h	2	140	140	104	102	147	147
Pop2	h	2	134	136	104	104	159	135
Pop2	h	2	134	140	nd	.	147	149
Pop2	h	2	−1	−1	104	104		
Pop3	d	1	134	134	104	104	147	147
Pop3	d	1	134	140			147	147
Pop3	d	1	134	134			145	147
Pop3	h	1	134	134	104	104	145	151
Pop3	h	1	134	136	104	106	147	145
Pop4	d	3	134	134	104	104	147	147
Pop4	d	3	136	136	104	98	147	147

Figure 1 Sample input file for MSA based on a two-column data matrix.

very large, the random discarding can result in significant variation in locus specific estimators. To account for this, MSA calculates heterozygosity, variance in repeat number (PCR product size) and number of alleles by averaging 200 randomly discarded data sets. This method provides significantly better estimates than a single randomly discarded data set. *F*-statistics can also be performed by the same procedure, but nondiscarded data sets are not biased and do not deviate from the mean of 1000 randomly discarded data sets (data not shown).

Data export. MSA provides input files for GENEPOP (Raymond & Rousset 1995), MSVAR (Beaumont 1999), STRUCTURE (Pritchard *et al*. 2000), ARLEQUIN (Schnei der *et al*. 1997) and MIGRATE (Beerli & Felsenstein 2001). Genetic distances are given as a PHYLIP (Felsenstein 1991) compatible distance matrix(es).

Error checking. MSA was intensively tested by comparison to the results of other software packages including MICROSAT (Minch *et al*. 1995), FSTAT (Goudet 1995), GENETIX (Belkhir *et al*. 1996–98) and ARLEQUIN (Schneider *et al*. 1997). No inconsistency was noted in these comparisons. Only the *P*-values of pairwise population comparisons differed between FSTAT (version 2.91) and MSA. Preliminary results suggest that FSTAT permutes among all available genotypes, but MSA and GENETIX only permute between the two groups for which the pairwise *F*-statistics are determined.

Availability. MSA was written in C/C++ and executables for Windows, MacOS9 and MacOSX can be downloaded from the authors' webpage (*http://i122server.vu-wien.ac.at/*).

Acknowledgements

Many thanks to B. Harr, M. Kauer and G. Muir for testing previous versions of the MSA software. Funding was provided by Fonds zur Förderung der wissenschaftlichen Forschung (FWF) grants and an EMBO Young Investigator Award to CS.

References

Beaumont MA (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the USA*, **98**, 4563–4568.

Belkhir K, Borsa P, Goudet J, Chikhi L, Bonhomme F (1996–98) *GENETIX, logiciel sous Windows™ pour la génétique des populations*. Laboratoire Génome et Populations, CNRS UPR 9060, Université de Montpellier II, Montpellier (France).

Ewens WJ (1972) The sampling theory of selectively neutral alleles. *Theoretical Popul Biology*, **3**, 87–112.

Felsenstein J (1991) *PHYLIP, Version 3.57c*. University of Washington, Seattle.

Goudet J (1995) FSTAT, Version 1.2.a program for IBM PC compatibles to calculate Weir and Cockerham's estimators of *F*-statistics. *Journal of Heredity*, **86**, 485–486.

Irvin SD, Wetterstrand KA, Hutter CM, Aquadro CF (1998) Genetic variation and differentiation at microsatellite loci in Drosophila simulans: evidence for founder effects in new world populations. *Genetics*, **150**, 777–790.

Kauer M, Zangerl B, Dieringer D, Schlötterer C (2002) Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. *Genetics*, **160**, 247–256.

Kimura M, Ohta T (1975) Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proceedings of the National Academy of Sciences of the USA*, **72**, 2761–2764.

Minch E, Ruiz-Linares A, Goldstein D, Feldman M, Cavalli-Sforza LL (1995) MICROSAT *Version 1.4d: a computer program for calculating various statistics on microsatellite allele data*. http://hpgl.stanford.edu/projects/microsat/

Payseur BA, Cutter AD, Nachman MW (2002) Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Molecular Biological Evology*, **19**, 1143–1153.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Schlötterer C (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics*, **160**, 753–763.

Schlötterer C, Vogl C, Tautz D (1997) Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. *Genetics*, **146**, 309–320.

Schneider S, Kueffer J-M, Roessli D, Excoffier L (1997) ARLEQUIN, *Version 1.1. A software for population genetic data analysis*. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.