Molecular Ecology Notes (2007) 7, 592–593 doi: 10.1111/j.1471-8286.2007.01784.x

PROGRAM NOTE

FORMATOMATIC: a program for converting diploid allelic data between common formats for population genetic analysis

NICHOLAS C. MANOUKIS*
Department of Ecology and Evolutionary Biology, University of California, Los Angeles, 621 Charles Young Dr So., Los Angeles, CA 90095, USA

Abstract

There has been a great increase in both the number of population genetic analysis programs and the size of data sets being studied with them. Since the file formats required by the most popular and useful programs are variable, automated reformatting or conversion between them is desirable. FORMATOMATIC is an easy to use program that can read allelic data files in GENEPOP, RAW (CSV) or CONVERT formats and create data files in nine formats: RAW (CSV), ARLEQUIN, GENEPOP, IMMANC/BAYESASS +, MIGRATE, NEWHYBRIDS, MSVAR, BAPS and STRUCTURE. Use of FORMATOMATIC should greatly reduce time spent reformatting data sets and avoid unnecessary errors.

Keywords: allele, conversion, data file, format, genotype, population genetic

Received 11 January 2007; revision accepted 6 March 2007

Population genetic analysis programs are becoming ever more sophisticated and specialized. At the same time, the size of data sets being analysed with them is also increasing, thanks to automation and high-throughput molecular techniques. This combination requires ever more manipulation of large data sets into several specialized formats to be used by a combination of programs for population genetic analysis. Illustrating the problem, a recent review by Excoffier & Heckel (2006) discusses over 20 computer programs that together provide a powerful pallet of analytical options, but which between them employ dozens of file types. Formatting a large gentoypic data set for use with these and other programs necessitates automated tools to avoid errors and time expenditure on repetitive tasks.

FORMATOMATIC is a simple program designed to automate the task of converting files containing diploid allelic data between formats needed for population genetic analysis. It does not, as a rule, produce files that can be used immediately by an analysis program. Usually, run options, metadata and other details required by a particular program and analysis must be set by the user. However, with FORMATOMATIC, I aim to reduce or eliminate the error prone and time-consuming task of re-formatting the actual data set. FORMATOMATIC complements the functionality offered by another specialized file conversion tool, CONVERT (Glaubitz 2004) and the data-importation abilities of some analysis programs.

Three specific advantages of FORMATOMATIC are worth highlighting: (i) the program is cross-platform, written in the Java programming language. It runs on any computer that has the Java Virtual Machine installed, including versions of Microsoft Windows, Apple Mac OS X, Unix and GNU/Linux. (ii) FORMATOMATIC is very easy to use. It consists of a simple graphical user interface where in-file and out-file names are set and formats selected. (iii) The source code is available for inspection and modification under the terms of the GPL, or General Public License. This permits users to add and contribute functionality as well as audit exactly how their data are being reformatted.

As of this writing, FORMATOMATIC can convert from files in GENEPOP (Raymond & Rousset 1995), RAW (CSV) or CONVERT (Glaubitz 2004) formats to RAW (CSV), ARLEQUIN (Excoffier *et al*. 2005), GENEPOP, IMMANC/BAYESASS +

Correspondence: N. C. Manoukis, *Present address: Section of Vector Biology, Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 12735 Twinbrook Parkway, Rockville, MD 20852, USA. Fax: (413) 521 9175;
E-mail: manoukis@taylor0.biology.ucla.edu

© 2007 The Author
Journal compilation © 2007 Blackwell Publishing Ltd

(Rannala & Mountain 1997), MIGRATE (Beerli 2006), NEWHYBRIDS (Anderson & Thompson 2002), MSVAR (Beaumont 1999), BAPS (Corander *et al*. 2004) or STRUCTURE (Pritchard *et al*. 2000; Falush *et al*. 2003) files. Documentation included with the program download provides more details on running FORMATOMATIC and on the file formats it can accept or produce. The program and documentation may be downloaded from: http://taylor0.biology.ucla.edu/~manoukis/Pub_programs/Formatomatic.

Together with the aforementioned CONVERT and the import capability of some of the analysis programs, files in one of FORMATOMATIC's in-formats can be transformed for use with 15 of the 23 analysis programs highlighted by Excoffier & Heckel (2006) as essential. FORMATOMATIC specifically supports creation of files for IMMANC/BAYESASS +, MIGRATE, NEWHYBRIDS, MSVAR and BAPS, which are not currently supported by others. I plan to add further input or output formats as new analysis programs are released or as requested by users, when possible. Contributions to the source code are encouraged.

Acknowledgements

Thanks to C. Taylor for supporting for this work. Thank you also to D. Earl for testing and suggestions. This work was supported by National Institutes of Health grants 5R01AI051633 and 5R01AI040308.

References

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, **160**, 1217–1229.

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

Beerli P (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, **22**, 341–345.

Corander J, Waldmann P, Marttinen P, Sillanpaa MJ (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**, 2363–2369.

Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: a survival guide. *Nature Reviews Genetics*, **7**, 745–758.

Excoffier L, Laval G, Schneider S (2005) ARLEQUIN, version 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Glaubitz JC (2004) CONVERT: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Molecular Ecology Notes*, **4**, 309–310.

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

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences, USA*, **94**, 9197–9201.

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

© 2007 The Author
Journal compilation © 2007 Blackwell Publishing Ltd