Genotype transposer: automated genotype manipulation for linkage disequilibrium analysis

David G. Cox* and Federico Canzian

Genome Analysis Group, International Agency for Research on Cancer, 69008 Lyon, France

Received on March 26, 2001; revised and accepted on April 25, 2001

ABSTRACT

Summary: The purpose of this work is to provide the modern molecular geneticist with tools to perform more efficient and more accurate analysis of the genotype data they produce. By using Microsoft Excel macros written in Visual Basic, we can translate genotype data into a form readable by the versatile software ‘Arlequin’, read the Arlequin output, calculate statistics of linkage disequilibrium, and put the results in a format for viewing with the software ‘GOLD’.

Availability: The software is available by FTP at: ftp://xcsg.iarc.fr/cox/Genotype_Transposer/.

Contact: cox@iarc.fr

Supplementary information: Detailed instruction and examples are available at: ftp://xcsg.iarc.fr/cox/Genotype_Transposer/. Arlequin is available at: http://lgb.unige.ch/arlequin/. GOLD is available at: http://www.well.ox.ac.uk/asthma/GOLD/.

Coupled with the increases in knowledge of human genetic diversity and the explosion of high-throughput genotyping methods, molecular geneticists are faced with a staggering amount of data to handle and analyze. Automated methods of data management and analysis are critical to rapidly and accurately reach conclusions with genotypic data. It is with this in mind that we have written ‘Genotype Transposer’.

Genotype transposer is a suite of Microsoft Excel macros, written in Visual Basic. The software takes genotype data from Single Nucleotide Polymorphisms (SNPs), in the form of an Excel worksheet, and organizes it into separate worksheets for each polymorphism. Then the software calculates the allele and genotype frequencies of each polymorphism, and checks if the alleles are in Hardy–Weinberg equilibrium. A second macro can then transform the genotype data into a format suitable for entry into the software ‘Arlequin’ (Schneider et al., 2000). This software, estimated to be in use by over 3000 scientists, can calculate the haplotype frequencies of the population in question, using the Expectation–Maximization algorithm. After the user runs the data through Arlequin, macros can extract the multi- and bi-locus haplotypes from the Arlequin output. Arlequin will not give D and D' statistics from genotypic data with unknown phase. However, we can use the bi-locus haplotypes as ‘observed’, and calculate expected haplotype frequencies from the allele frequencies of the two polymorphisms. With this information, we can calculate D and D', popular coefficients of linkage disequilibrium (Devlin and Risch, 1995). The significance of the difference of D' from zero can also be calculated (Elbein, 1992). The last function of the macro is to perform these calculations, and transform them into a format suitable for entry into the ‘GOLD’ software (Abecasis and Cookson, 2000). This software displays linkage statistics in an easy to interpret graph.

See Figure 1 for a schematic diagram of data flow within genotype transposer. Multiple ‘genotype transposer’ toolbar items can occur on the toolbar if the user opens more than one worksheet with the macro in it. This should not be a problem, but the user should be aware that the macro is run on the data in the current window. The user could run all macros in the suite at once, as they are executed.

Genotype Data Entered

Data Organized

Arlequin input made

Arlequin calculates

haplotype freqs.

Multi-locus haplotypes extracted + displayed

Bi-locus haplotypes extracted + displayed

Bi-locus haplotypes extracted

GOLD input made

Fig. 1. Schematic diagram of data-flow using the genotype transposer macro.

*To whom correspondence should be addressed.
Genotype transposer

sequentially. After the macro that creates the form for input into Arlequin is run, the user will be presented with a button to click. Before clicking this button, the user can open their Arlequin application, and run the analysis. After the analysis is finished, all that needs to be done is to click the ‘OK’ button, and the macro will extract the data from the Arlequin output. Unfortunately, missing data cannot be included in Arlequin input, as this causes Arlequin to see missing data as another allele. Also, the macros will only work with SNP data. For more detailed information, please read the README.txt and one of the instructions files.

REFERENCES


