COMPUTER PROGRAMS

# GeneticStudio: a suite of programs for spatial analysis of genetic-marker data

RODNEY J. DYER
*Department of Biology, Virginia Commonwealth University, Richmond, Virginia 23284-2012, USA*

## Abstract

**The analysis of genetic marker data is increasingly being conducted in the context of the spatial arrangement of strata (e.g. populations) necessitating a more flexible set of analysis tools. GeneticStudio consists of four interacting programs: (i) Geno a spreadsheet-like interface for the analysis of spatially explicit marker-based genetic variation; (ii) Graph software for the analysis of Population Graph and network topologies, (iii) Manteller, a general purpose for matrix analysis program; and (iv) SNPFinder, a program for identifying single nucleotide polymorphisms. The GeneticStudio suite is available as source code as well as binaries for OSX and Windows and is distributed under the GNU General Public License.**

*Keywords*: genetic structure, population graphs, spatial structure

*Received 18 August 2008; revision accepted 22 August 2008*

The analysis of genetic marker data is increasingly being conducted in the context of the spatial arrangement of strata (e.g. sites or populations) as well as the importance of intervening geographical features. The addition of spatial data to population genetic analysis requires flexible tools that are designed to incorporate nongenetic and nonstrata data beyond simple pairwise distance analysis. The analysis of spatial genetic data also needs powerful visualization tools to aid in the interpretation of increasingly complicated and precise studies. GeneticStudio is a suite of programs that have been developed to analyse genetic marker data with particular emphasis on the inclusion of spatial covariate data. This covariate data can be GIS locations as well as any other ecological or local variable measured on individuals or as a property of a stratum (e.g. a population location). These programs were designed with particular attention to the ease of data input, intuitive user-interfaces, interapplication data exchange, and cross-platform deployment. All programs accept tab-delimited text files as both input and export similar files. At present, GeneticStudio consists of four programs (Geno, Graph, Manteller, SNPFinder) and a comprehensive user's manual and example data sets. What follows is a brief outline of the basic functionality of each component program. For a more in-depth treatment of the capabilities of GeneticStudio, download the software package and/or the Users Manual from http://dyerlab.bio.vcu.edu.

*Geno:* This is the main program for interacting with individuals, their genotypes, stratum and covariates. Data are imported into Geno from tab-delimited text files (as from Excel) via a set of import dialogs that allow users to quickly characterize all data types in their input data set. Supplemental data from external text files (including additional loci, strata, and covariates) can be merged into existing data sets. Data types accepted into Geno fall into the following three categories. *Stratum* variables denote populations, regions, or any other arbitrary partitioning of individuals that is categorical. *Covariate* data variables are treated as continuous variables and can represent spatial, ecological, or any other data that is measured at the level of the individual (although not necessarily unique to each individual). Genetic *marker* data types are classified as either *binary genetic markers* (e.g. amplified fragment length polymorphisms, intersimple sequence repeats) representing alleles that are present or absent, *haploid genetic markers* (e.g. chloroplast microsatellites, haploid species), and *codominant genetic markers* [allozymes, microsatellites, single nucleotide polymorphisms (SNPs)]. Geno has the ability to import genotypes with ploidy up to tetraploids (beyond if no editing of the individual genotypes is required) as well as

Correspondence: Rodney J. Dyer, Fax: (804) 828-0874; E-mail: rjdyer@vcu.edu

**Table 1** Analysis available from within Geno and associated data types that are used. Data type abbreviations are: strata (*Z*; categorical data such as population or regions), covariate data (*S*; spatial and/or ecological variables measured at the individual), and genetic data as dominant loci (*D*; e.g. AFLP's, ISSR's), haploid loci (*H*; SNP's, organelle microsatellites) and codominant loci (*C*; allozymes, microsatelites, EPICs, SNPs)

| Category | Routine | Data type | Reference |
|---|---|---|---|
| Distances | AMOVA distance | *D H C* | Excoffier *et al.* (1992); Smouse & Peakall (1999) |
| | Genetic covariance† | *D H C* | Dyer *et al.* (2004) |
| | Relatedness‡ | *C* | Lynch & Ritland (1999) |
| | Covariate§ | *S* | |
| | Hypothesis¶ | *Z* | Johnson & Wichern (1992) |
| | Bin distances** | *Z S* | Smouse & Peakall (1999) |
| Diversity | PCA | *Z D H C* | Westfall & Conkle (1992) |
| | Heteroscedasticity | *Z D H C* | R.J. Dyer (unpublished) |
| | Summary stats†† | *Z D H C* | Hedrick (2005) |
| | Diversity gradient | *Z C S* | |
| Frequencies | Tables‡‡ | *Z D H C* | |
| | Frequency gradient | *Z D H C S* | |
| | Pies on map§§ | *Z D H C S* | |
| Structure | *F*-statistics | *Z C* | Hedrick (2005) |
| | AMOVA | *Z D H C* | Excoffier *et al.* (1992) |
| | StAMOVA | *Z D H C S* | Dyer *et al.* (2004) |
| | Population Graph¶¶ | *Z D H C* | Dyer & Nason (2004) |
| Matrix | Mantel test | *Z D H C S* | Mantel (1967) |

†following Gower (1966); ‡calculated symmetrically; §distances are calculated as Euclidean distance between multivariate covariates as either individual distances or as distances among strata; ¶creates idempotent hypothesis (**H**) matrix; **strata and spatial distances are used to create bin distance matrices ($\mathbf{X}^{(n)}$) for spatial autocorrelation; ††includes A, $A_e$, $A_{95\%}$, $H_O$, $H_E$, $H_E$*; ‡‡can be calculated in total or along any arbitrary stratum variable; §§requires Google Maps or GoogleEarth (available for free from http://www.google.com); ¶¶requires R (http://cran.r-project.org) to be installed locally.

data from mixed haploid/diploid systems. Each genetic data type can be edited within Geno using custom spreadsheet widgets (e.g. an editing field for each allele within a column of genotypes for the individual).

Geno is particularly well suited to manipulating large data sets by allowing the identification of arbitrary subsets of data by selecting from existing strata, covariates, and genetic loci. The limits to the amount of data that can be worked with are dependent solely upon the amount of memory in the computer used. Strata can be exported from Geno as KML files to create maps using GIS software such as ArcGIS (http://www.esri.com) or GRASS GIS (http://grass.itc.it/) or mapping tools such as GoogleEarth or Google Maps. Genotypes and associated stratum and covariate data can also be exported from Geno as multivariate data sets (following Westfall & Conkle 1992; Dyer *et al.* 2004) for analysis in external programs such as R or SAS.

Genetic, strata, and covariate data can be easily translated into matrices and analysed within Geno or exported to external programs (such as Manteller) for subsequent analysis. Genetic conversion includes distance calculations based upon genetic metrics at individual and strata levels, and pairwise relatedness (Table 1). A combination of strata and spatial data can be used to create hypothesis matrices at defined distance bins for spatial autocorrelation analysis (after Smouse & Peakall 1999). Distances based upon spatial or ecological covariates are calculated based upon multivariate Euclidean distances. Where appropriate, intermediate matrices used in the execution of analysis are made available as output options allowing subsequent analysis as well as practitioners to understand intermediate steps in the analysis. When matrices are estimated, summary statistics and distributions can also be provided, however, more flexibility in analysing matrix structure and performing matrix analysis is gained by exporting matrices from Geno into Manteller.

Genetic analyses are partitioned into four categories (Table 1). Diversity statistics include general summary statistics, tests for unequal within-population diversity (heteroscedasticity), principal components analysis, and diversity statistics by strata as a function of spatial variables. Frequency analysis include tabulation of allele frequencies in total or by arbitrary stratum, frequencies as a function of covariate (spatial) variables, and the ability to export frequency pie charts to GoogleEarth. Analysis based upon genetic structure include hierarchical *F*-statistics (Wright 1978), AMOVA (Excoffier *et al.* 1992), StAMOVA (Dyer *et al.* 2004), and Population Graphs (Dyer & Nason 2004; Dyer 2007).

Matrix analysis allows you to perform isolation by distance calculations based upon any of the matrix distance metrics submitted to the Mantel test.

All results in Geno are presented as HTML in the results pane and can be printed or exported as PDF documents. Where appropriate, results are presented in both tabular and graphical formats to aid in data exploration. Items can be copied from the Results pane and pasted into other program on all platforms. Geno saves both data and results in a single platform-independent binary file whose format is described in source code.

*Graph:* This program allows the visualization and analysis of network topologies in two-dimensional space. While this program is primarily focused on the analysis of Population Graph structure (Dyer & Nason 2004; Dyer 2007) it is amenable to analysis of any nondirectional graph topologies. Population Graph input files are created in Geno or via the online analysis from http://dyerlab.bio.vcu.edu. The spatial arrangement of graph topologies is determined from a modified spring model and can be directly manipulated by the user. Node, edge, and background colours and label fonts are all customizable to produce presentation-quality images.

Spatial data can be inserted into graph topologies allowing two spatially explicit topological analyses. First, the distance among nodes within the topology and their spatial distances facilitates the analysis of isolation by graph distance (IBGD; Dyer & Nason 2004; R.J. Dyer *et al*. in preparation). IBGD has particularly favourable statistical properties including homoscedasticity and stability as estimation of all pair-wise between populations elements are performed using the entire data set rather than individual pairs of populations. Significance of IBGD patterns is ascertained using a Mantel Test. Second, assigning spatial locations to individual nodes allows the identification of compressed and extended edges. Compressed edges are populations whose spatial location is closer than would be expected given their genetic covariance whereas extended ones are those who are spatially further than expected (R.J. Dyer *et al*., in preparation). The node set as well as sets for compressed, normal, and extended edges can be visualized 'on the ground' by exporting the topology as a KML file amenable for viewing in ArcGIS, GRASS GIS, or GoogleEarth.

Congruence graphs, or the topology that results from the intersection of two graphs that have identical node sets but potentially different edge sets, can be estimated from within Graph. The probability of observing a congruence graph with an edge set as large as the observed one is given using a combinatorial approach. Once imported into Graph, congruence graphs can be analysed just as any other graph topology.

Features within a graph topology can be analysed as matrices, distance measures, and centrality parameters. Matrices include adjacency and binary adjacency matrices. Distance analyses include the estimation of shortest path topology and the analysis of IBGD. Finally, centrality parameters include degree, clique, and eigenvalue centrality. All results are presented in a results pane as in Geno and can be either printed or saved as PDF documents.

*Manteller:* This is a general matrix manipulation and analysis program. Input matrices for Manteller consist of tab-delimited text files as exported from both Geno and Graph as well as from external sources (e.g. Excel, R, SAS). Any number of matrices may be imported into Manteller for analysis and matrices can be exported from Manteller as text files for external analysis. Once imported into Manteller, all elements of the matrices can be edited directly within the GUI and the size of the matrix that can be edited and operated on is only limited by the amount of computer memory (e.g. there is no arbitrary limit to number of rows or columns).

Manteller allows all matrix operations, conversions, and analysis. Matrix operations include transpose, generalized inverse, normalization, Hadamard products, matrix scaling, addition and matrix multiplication. Matrix conversions include binarizing a matrix, translating between distance, covariance, and correlation matrices. Analysis include general summarizing of matrix components (partitioned by diagonal and off-diagonal), matrix pair scatter plots, and Mantel tests. Multiple Mantel tests are also available in a stepwise fashion allowing the user to create arbitrarily large models.

*SNPFinder:* This program is useful in the discovery of SNPs from methods such as Solexa sequencing. SNPFinder inputs a standard FASTA file and saves individual sequences, sequence tallies, and putative SNP loci as tab-delimited text files.

The main functionality of SNPFinder lies in filtering large sets of sequences. Filtering operations include collapsing the set of all sequences to those that are unique (e.g. tallying multiple sequences), selecting sequences that occur at particular frequencies, including or excluding sequences that have particular motifs as defined by fixed strings or Regular Expressions, and creating the reverse compliments. Once a filtered set of sequences are identified, individual SNPs or SNP families can be identified. Putative SNPs with relative frequencies of each variant can be saved as text for subsequent analysis.

All software programs are available as source-code (a mixture of C and C++) as well as precompiled binaries for both Windows and Macintosh platforms and can be freely downloaded from the software page at http://dyerlab.bio.vcu.edu. All software and documentation are licensed under the GNU GPL (version 3; http://www.gnu.org/licenses/gpl.html). Non-English native language translations (e.g. all text in menus and dialogs and keyboard shortcuts appropriate to users' operating system) are currently being developed and will be distributed with each release. Persons interested in aiding in development or translation of GeneticStudio should contact the author.

## Acknowledgements

I would like to thank J.F. Fernandez, B.L. Brown, R.C. Garrick, J.D. Nason and the members of the Dyerlab for extensive testing of this software. This work was partially supported by two grants from the National Science Foundation (DEB-053102 and DEB-0640803) to R.J.D.

## References

Dyer RJ (2007) The evolution of genetic topologies theoretical. *Population Biology*, **71**, 71–79.

Dyer RJ, Nason JD (2004) Population Graphs: the graph theoretic shape of genetic structure. *Molecular Ecology*, **13**, 1713–1728.

Dyer RJ, Westfall RD, Sork VL, Smouse PE (2004) Two-generation analysis of pollen flow across a landscape V: a stepwise approach for extracting factors contributing to pollen structure. *Heredity*, **92**, 204–211.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Hedrick PW (2005) *Genetics of Populations*, 3rd edn. Jones and Bartlett Publishers, Sudbury, Massachusetts.

Johnson RA, Wichern DW (1992) *Applied multivariate statistical analysis*. Third edition. Prentice Hall, New Jersey.

Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. *Genetics*, **152**, 1753–1766.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Smouse PE, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, **82**, 561–573.

Westfall RD, Conkle MT (1992) Allozyme markers in breeding zone designations. *New Forests*, **6**, 279–309.

Wright S (1978) *Evolution and the Genetics of Populations; Volume 4: Variability Within and Among Natural Populations*. University of Chicago Press, Chicago.