Molecular Ecology Notes (2005) doi: 10.1111/j.1471-8286.2005.01113.x

PROGRAM NOTE

GENER: a server-based analysis of pollen pool structure

RODNEY J. DYER
Department of Biology, Center for the Study of Biological Complexity, Virginia Commonwealth University, 1000 West Cary St., Richmond, Virginia 23284-2012 USA

Abstract

The server-based program GENER performs the two-generation analysis of pollen flow for data consisting of mother/offspring arrays using genetic markers. The GENER program decomposes the genetic variance sampled by maternal individuals within and among pollen pool components of genetic variance and is accessible from http://dyerlab.bio.vcu.edu. These estimates are used to construct the test statistic, Φ_{FT}, whose significance is tested via permutation. The Φ_{FT} statistic can subsequently be used to quantify genetic effective pollen donor population size (N_{ep}), effective mating area and dispersal distance. Furthermore, the GENER program can calculate Φ_{FT} values for all pairs of substrata within the data set.

Keywords: gene flow, pollen structure, two-generation analysis

Received 13 April 2005; revision accepted 15 June 2005

Pollen-mediated gene movement plays a critical role in the evolutionary dynamics of plant populations (e.g. Harper 1977). The analysis pollen movement has become increasingly prevalent among population and conservation geneticists as they seek to understand how genes move across a landscape and what site-specific ecological factor influences this movement (e.g. Sork *et al*. 1998, 1999). A recent addition to the repertoire of tools aimed at quantifying contemporary patterns of pollen movement is the 'two-generation' approach presented by Smouse *et al*. (2001). This analytical framework is based on the AMOVA approach from Excoffier *et al*. (1992) but quantifies the differences among sampled pollen pools rather than populations of adults. Under the two-generation model, spatially separated maternal individuals act as 'pollen samplers'. The differences in pollen haplotypes sampled across the landscape provide estimates of average pollen dispersal distance, effective mating area and genetic effective adult pollination population size (e.g. Austerlitz & Smouse 2001).

The GENER program introduced herein performs the decomposition of genetic variance following the linear model approach proposed by Dyer *et al*. (2004). This linear model approach, which Dyer *et al*. termed STAMOVA, is a multivariate translation of the AMOVA model from Excoffier *et al*. (1992). In fact, the AMOVA model is simply a special case of the more general STAMOVA family of linear modes. These linear models partition the total multilocus genetic variance, σ^2_T, into the additive components representing the genetic variance within the maternally sampled pollen pools, σ^2_W, and the genetic variance among maternally sampled pollen pools, σ^2_A. Once estimated, these quantities yield the statistic, Φ_{FT}, which is a standardized measure of how much of the total genetic variance is due to differences among sampled pollen pools, relative to all the pollen haplotypes examined. This statistic is identical to the Φ_{ST} statistic estimated from AMOVA; the only difference being that Φ_{ST} measures differences among strata (the *S*) and Φ_{FT} measures differences among families (the *F*; Smouse *et al*. 2001).

This GENER program accepts genetic marker data from codominant (e.g. allozymes and microsatellites) or dominant [e.g. amplified fragment length polymorphisms (AFLPs) and restriction fragment length polymorphisms (RFLPs)] genetic markers. The input file format for this program and all other programs on the same cluster are tab-delimited text files. These files are essentially what you get from your spreadsheet program when you save the file as text. There is no special formatting required. Allozyme genotypes are denoted by pairs of numbers zero to nine. Microsatellite genotypes can be encoded in three different formats: (i) two columns of data for each locus; (ii) each allele separated by a period (e.g. 123.87) or (iii) a fixed width representation (e.g. as 123087 using the same alleles as the last example). Binary markers are encoded using 0/1.

Missing data are handled textually, by encoding each missing allele as a negative number. By convention, the

Correspondence: R. J. Dyer, Fax: (804)828-0503; E-mail: rjdyer@vcu.edu

number –9 is used, although any number strictly less than zero is acceptable. The zero allele, 0, is treated as a functioning allele. In the analysis, the GENER program uses a posterior likelihood method for dealing with missing data. The program treats a missing genotype as all possible genotypes, with probability equal to their posterior probability existing in the data set. This is a conservative approach, favouring the null hypothesis of no differences among sampled pollen pools.

Data files for the GENER program consist of two categories of individuals from which you have extracted genotypes; mothers and offspring. As such, there is a requirement that you should include the genotypes of both sets of individuals in a single data file. I have devised the following scheme to differentiate among maternal and offspring genotypes. First, each maternal individual has a unique identification, referred to as the 'maternal identification number' or MIN. A MIN can consist of any combination of numbers or letters, even including spaces. The only restriction on the content of a MIN is that it cannot include tab or comma characters as they are used to distinguish separate columns of data in your data file. All offspring from the same mother will have the same MIN. Next, all offspring will have what I refer to as an 'offspring identification number' or OIN. The OIN for half-sib offspring (e.g. those that share the same maternal individual) can be numbered however you like. However, the OIN for the maternal individual *must* be zero (0). This is the only restriction. An OIN = 0 tells the program that these genotypes belong to a maternal individual, whereas any other number in the OIN designates that individual as an offspring.

Significance testing is conducted via permutation following Smouse *et al*. (2001), and proceeds as follows. First, the null hypothesis, H_O, states that there is no significant difference among the pollen pools being assayed. If the null hypothesis is true, then the multilocus pollen genotype is equally likely to be sampled from any of the potential maternal individuals. In essence, the null hypothesis states that the allocation of genotype to strata is irrelevant. As such, we may permute the pollen genotypes among maternal strata and create a null distribution of σ_A^2. The GENER program does this 999 times and includes the original σ_A^2 in the calculation of the probability of observing a value of σ_A^2 equal to or larger than the observed value. The GENER program provides these data in summary form and as a textual bar chart.

The GENER program can be found on the software page on the Dyer Laboratory web server (http://dyerlab.bio.vcu.edu). The program is run directly on the server and depending upon the size of your data set and the current load on the server, the results may either be returned to you directly, or emailed to you in either textual or pdf format. Documentation and sample data sets for this program and all other programs hosted on these servers are available on the same web page as the program links. Prior to running GENER, you provide the following information via web page interface:

(1) Your email address in case the load is too large and the results must be emailed to you. The amount of time a web browser will wait on a web page to load is variable and I found it beneficial to simply email results if the analysis will take over *c*. 20 s. Currently, simultaneous significance testing of two data sets using 1000 permutations of a data set with 600 individuals and 30 polymorphic loci takes roughly 20 s to complete on the server.

(2) The data file. The actual format of the data file is discussed earlier. In addition, you will include information on the column separators (tab or comma are supported) and whether the data have a header row or not.

(3) Genetic data format. This information includes the marker being used (allozyme, microsatellite or binary) as well as the identification of which columns in your data set represent the MIN, the OIN and at which column the genotypes start.

(4) Output parameters. These selections determine if you want to test significance, perform pairwise analyses of all strata, and whether the output should be in text, html or pdf.

Once you define these parameters, the server will upload your data file and analyse it. Once the analysis is completed and you are presented with the results, your data are deleted from the server. The GENER program is not a data repository, if you wish to perform subsequent analyses on the same data set, you will have to provide it each time.

The GENER program has been implemented in Objective-C and runs on a dedicated Apple XServe Cluster in the Dyer Laboratory (http://dyerlab.bio.vcu.edu). The GENER program is deployed as a server-based analysis rather than a stand-alone application for three reasons. First, dedicated servers typically have more computational resources at their disposal than the average laptop or desktop machines. For example, the primary server supporting the GENER program is a dual processor G5 with 4 GB of RAM. Second, by using a web interface, all computational platforms are supported. As long as you have access to the Internet, you can run your analyses. Finally, since there is only one executable, any updates or bug fixes are immediately distributed to all potential users.

References

Austerlitz F, Smouse PE (2001) Two-generation analysis of pollen flow across a landscape II: relation between pollen dispersal and interfemale distance. *Genetics*, **157**, 851–857.

Dyer RJ, Westfall RD, Sork VL, Smouse PE (2004) Two-generation analysis of pollen flow across a landscape IV: 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 inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Harper JL (1977) *Population Biology of Plants*. Academic Press, Orlando Florida.

Smouse PE, Dyer RJ, Westfall RD, Sork VL (2001) Two-generation analysis of pollen flow across a landscape I: male gamete heterogeneity among females. *Evolution*, **55**, 260–271.

Sork VL, Campbell D, Dyer RJ *et al.* (1998) *Proceedings from A Workshop on Gene Flow in Fragmented, Managed, and Continuous Populations*. National Center for Ecological Analysis and Synthesis, Santa Barbara, California. http://www.nceas.ucsb.edu/papers/geneflow.

Sork VL, Campbell D, Nason J, Fernandez JF (1999) Landscape approaches to historical and contemporary gene flow in plants. *Trends in Ecology and Evolution*, **14**, 219–223

