SMOGD: USER MANUAL

Table of Contents

Introduction

Purpose

SMOGD (Software for the Measurement of Genetic Diversity) is a web based application for the calculation of genetic diversity. Specifically, it calculates $G_{ST\ est}$ (Nei, 1983) G'_{ST} (Hedrick, 2005) and D_{est} (Jost, 2008). It also generates bootstrap replicates of data sets and uses these replicates to estimate standard error and variance of the aforementioned parameters.

Parameters Calculated:

Basic Parameters: assumes actual allele frequencies are known

- **n** = number of populations
- $\mathbf{D_{ST}}$ = absolute differentiation (Nei, 1973)
- $\mathbf{G_{ST}}$ = relative differentiation (Nei, 1973)
- $\mathbf{H_{ST}}$ = between-subpopulation heterozygosity (Aczel & Daroczy, 1975; Tsallis & Brigatti, 2004)
- $\mathbf{\Delta_{ST}}$ = between-subpopulation component of diversity, or the effective number of distinct subpopulations (Jost, 2008)
- **D** = actual differentiation (Jost, 2008)
- $\mathbf{H_S/H_T}$ = proportion intra-population heterozygosity vs total heterozygosity (Jost, 2008)
- $\mathbf{\Delta_S/\Delta_T}$ = proportion of total diversity that is contained in the average subpopulation (Jost, 2008)

Estimated Parameters: diversity measures for small sample sizes

- $\mathbf{\tilde{N}}$ = harmonic mean of population sizes
- $\mathbf{H_{S\ est}}$ = nearly unbiased estimator of within-subpopulation heterozygosty (Nei & Chesser, 1983)
- $\mathbf{H_{T\ est}}$ = nearly unbiased estimator of total-subpopulation heterozygosty (Nei & Chesser, 1983)
- $\mathbf{H_{ST\ est}}$ = nearly unbiased estimator of between-subpopulation heterozygosity (Nei & Chesser, 1983)
- $\mathbf{G_{ST\ est}}$ = nearly unbiased estimator of relative differentiation (Nei & Chesser, 1983)
- $\mathbf{G'_{ST\ est}}$ = standardized measure of genetic differentiation (Hedrick, 2005)
- $\mathbf{D_{est}}$ = estimator of actual differentiation (Jost, 2008)

Bootstrap Parameters

- Bootstrapped estimates of $G_{ST\ est}$, $G'_{ST\ est}$, and D_{est}, (= values of diversity indices averaged across replicates).
- Variance and standard error of the mean calculated from bootstrap replicates.

Distance Matrices

- Tables of pairwise distances for $G_{ST\ est}$, $G'_{ST\ est}$, and D_{est} for each locus.

File Formats

Import

SMOGD will import files in the GenePop (Raymond & Rousset,1995) and Arlequin (Excoffier et al., 1997). If your data is not in one of these formats I recommend using GenAlEx, an MS-Excel plugin for population genetic analysis to manipulate your data and export it in any one of the preceding formats.

GenePop

Details concerning GenePop format may be found at http://genepop.curtin.edu.au/

Arlequin

Details concerning Arlequin format may be found at http://cmpg.unibe.ch/software/arlequin3/

Export

Data is exported as both html and tab delimited files suitable for import into MS-Excel or database programs. The tab delimited files are time-stamped and the html links to these files are dynamically updated so that a user can only download files relating to the results of the data they submitted. Result files are deleted from the web-sever every 24 hours. Data sets are never saved although technically they exist in memory (RAM) until the user navigates away from the webpage.

Background

Basic and Estimated Parameters

SMOGD essentially calculates two sets of parameters. The 'basic parameters' correspond to the diversity measures reported in Table 1 of Jost (2008). They are presented as illustrative examples of how the parameters differ from each other. For actual data sets, where you have genotypes of individuals sampled from larger populations, the 'estimated parameters' more accurately account for small population sizes and associated sampling errors (Nei & Chesser, 1983).

Bootstrapping and Distance Matrices

Bootstrapping provides a way to estimate variance and standard error of the mean. Bootstrapping of subdivided population genetic data can be done at the population level (resampling populations) and the individual level (resampling individuals only) and at the individual and population level (resampling both populations and individuals). The implementation of the bootstrapping algorithm employed by SMOGD resamples at the individual level.

Generally, bootstrapping to estimate parameters (e.g., averaging $G_{st\ est}$ or D_{est} across replicates) does not provide good measures of diversity (Petit & Pons, 1998). However, it can be used to estimate variance and standard deviation of the mean (Jost, 2008; Chao et al., 2008). The recommendation then is: *don't report the bootstrap estimates of the estimated parameters, rather report the estimated parameters and the bootstrapped estimates of variance or standard deviation of the mean.*

The distance matrices are pairwise comparisons of populations on a locus by locus basis. Matrices are provided for $G_{st\ est}$, $G'_{st\ est}$, and D_{est}.

How to Cite

Crawford NG. 2009. SMOGD: Software for the Measurement of Genetic Diversity. Molecular Ecology Resources. In Prep.

Usage

The website is pretty self-explanatory. But, briefly: delete the sample data (control-A delete), paste in your file, select the number of bootstrap replicates (max = 1000), and click submit. When the analysis finishes the page will refresh with html output and links for downloading tables.

Known Bugs

As of 05/09

- **" Internal Server Error...":** This may occur for a number of reasons. Most likely you have managed to generate a 'divide by zero' error. Computers can't calculate '1/0' although Scipy is pretty robust to them (see below). If your file works without bootstrapping, while bootstrapping induces the error, check to see if any of the 'estimated parameters' are zero. If they are, you've found your problem.

  If you don't see any zeros in the 'estimated parameters' you may have found a bug. Shoot me an email and we'll figure it out.

- **'nan'** appears in results tables. Scipy is robust to 'divide by zero' type errors and reports them as 'nan.' It's possible to get divide by zero errors if populations are genetically similar. Bootstrapping may result in populations with no diversity especially if your populations are not particularly genetically diverse to begin with. See above.

- There have been a few reports of SMOGD hanging or crashing with large data sets. However, I've successfully ran data sets with 600 individuals, 10 loci, and 20 populations.

Citations

Aczél J, Daróczy Z. 1975. On measures of information and their characterizations. Mathematics in Science and Engineering, vol. 115, Academic Press, New York, San Francisco, London, 1975, xii + 234 pp.

Chao A, Jost L, Chiang SC, Jiang YH, Chazdon, RL. 2008. A two-stage probabilistic approach to multiple-community similarity indices. Biometrics, 64(4),1178-86

Excoffier L, Laval G, Schneider S. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50.

Hedrick, PW. 2005. A Standardized Genetic Differentiation Measure. Evolution 59(8), 1633-1638

Jost L. 2008. G_{ST} and its relatives do not measure differentiation. Molecular Ecology 17(18), 4015-4026

Nei M. 1973. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences, USA., 70(12, Pt 1), 3321-3323

Nei M, Chesser RK. 1983. Estimation of fixation indices and gene diversities. Annals of Human Genetics, 47, 253-259.

Petit RJ, Pons O. 1998. Bootstrap variance of diversity and differentiation estimators in a subdivided population. Heredity, 80(1), 56.

Raymond M, Rousset F. 1995. GENEPOP: population genetics software for exact tests and ecumenicism. Journal of Heredity, 86, 248-249.

Tsallis C, Bigatti E. 2004. Nonextensive statistical mechanics: A brief introduction. Continuum Mechanics and Thermodynamics, 16(3), 223-235

Acknowledgements

I would like to thank the following people: Karen Mock, Eric O'Neil, Mark Coulson, Naomi Dyer, and Niklas Tysklind who helped with debugging. Andy Reinmann who brainstormed the name 'SMOGD.' And, as always, Corinne Crawford provided inspiration.