

USER MANUAL vs 2.6: Software for the Measurement of Genetic Diversity (SMOGD)

Nicholas Crawford: ngcrawford@gmail.com

2009-09-29

1 Introduction

1.1 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) diversity indices. It also generates bootstrap replicates of data sets and uses these replicates to estimate standard error, variance, and 95% confidence intervals.

1.1.1 A Brief Note about D_{est} and $G'_{ST\ est}$

Although $G'_{ST\ est}$, like D_{est} , better accounts for populations with high allelic diversity, only D_{est} accounts for populations alleles alternatively fixed in different populations. To clarify: in a scenario where one sub-population is fixed at allele A, but ten other sub-populations are fixed at allele B, $G'_{ST\ est}$ will over report a population-wide diversity of 1.0 while D_{est} will report a population diversity of 0.55.

1.2 Parameters Calculated

1.2.1 Basic Parameters: assumes actual allele frequencies are known

- n = number of populations
- D_{ST} = absolute differentiation (Nei, 1973)

Equation: $Ht - Hs$

- G_{ST} = relative differentiation (Nei, 1973)

Equation: Dst/Ht

- H_{ST} = between-subpopulation heterozygosity (Aczel & Daroczy, 1975; Tsallis & Brigatti, 2004)

Equation: $\frac{(Ht-Hs)}{(1-Hs)}$

- Δ_{ST} = between-subpopulation component of diversity, or the effective number of distinct subpopulations (Jost, 2008)

Equation: $\frac{(1-Hs)^{-1}}{(1-Ht)^{-1}}$

- D = actual differentiation (Jost, 2008)

Equation: $(\frac{(Ht-Hs)}{(1-Hs)}) \cdot (\frac{n}{(n-1)})$

- H_S/H_T = proportion intra-population heterozygosity vs total heterozygosity (Jost, 2008)
- Δ_S/Δ_T = proportion of total diversity that is contained in the average subpopulation (Jost, 2008)

Equation: $\frac{(1-H_t)^{-1}}{(1-H_s)^{-1}}$

1.2.2 Estimated Parameters: diversity measures for small sample sizes

- \tilde{N} = harmonic mean of population sizes

Equation: $\frac{n}{\sum \frac{1}{\text{population size}}}$

- \tilde{N} **approximate** = approximation of harmonic mean as suggested by Anne Chao. It can handle negative numbers and thus is use for calculating the harmonic mean of D_{est} values which can be negative.

Equation: $\frac{1}{(1/\text{mean}) + \text{variance}(D_{\text{est}})(1/\text{mean})^3}$

- $H_{S \text{ est}}$ = nearly unbiased estimator of within-subpopulation heterozygosity (Nei & Chesser, 1983)

Equation: $\left(\frac{2 \cdot \tilde{N}}{2 \cdot \tilde{N} - 1}\right) \cdot H_s$

- $H_{T \text{ est}}$ = nearly unbiased estimator of total-subpopulation heterozygosity (Nei & Chesser, 1983)

Equation: $H_t + \frac{H_{S \text{ est}}}{(2 \cdot \tilde{N} \cdot n)}$

- $H_{ST \text{ est}}$ = nearly unbiased estimator of between-subpopulation heterozygosity (Nei & Chesser, 1983)

Equation: $\frac{(H_t - H_s)}{(1 - H_s)}$

- $G_{ST \text{ est}}$ = nearly unbiased estimator of relative differentiation (Nei & Chesser, 1983)

Equation: $\frac{(H_{t \text{ est}} - H_{s \text{ est}})}{H_{t \text{ est}}}$

- $G'_{ST \text{ est}}$ = standardized measure of genetic differentiation (Hedrick, 2005)

Equation: $\frac{(G_{st \text{ est}} \cdot (n-1 + H_{s \text{ est}}))}{((n-1) \cdot (1 - H_{s \text{ est}}))}$

- D_{est} = estimator of actual differentiation (Jost, 2008)

Equation: $\left(\frac{(H_{t \text{ est}} - H_{s \text{ est}})}{(1 - H_{s \text{ est}})}\right) \cdot \left(\frac{n}{(n-1)}\right)$

1.2.3 Bootstrap Parameters

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

Generally it is not appropriate to report the bootstrapped estimates of diversity. Rather, bootstrapping is included to provide measures of confidence.

1.2.4 Distance Matrices

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

2 File Formats

2.1 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.

2.1.1 GenePop

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

2.1.2 Arlequin

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

2.2 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-server every 24 hours. Data sets are never saved although technically they exist in memory (RAM) until the user navigates away from the webpage.

3 Background

3.1 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).

One common concern when interpreting the results is the presence of negative parameters. These occur when populations are almost identical/undifferentiated. In these situations, the results can be reported as 0. However, SMOGD does not make this conversion for you.

3.2 Bootstrapping and Distance Matrices

Bootstrapping provides a way to estimate variance, standard error of the mean, and 95% confidence intervals. 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, standard deviation of the mean, and 95% confidence intervals. I should also note that it is possible to calculate estimated values that fall outside of the range of the confidence intervals generated by bootstrapping. This seems to most commonly be a problem when a population/locus contains a very little diversity and has a very small value standard deviation.

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} . Distance matrices are also calculated across loci by taking the harmonic mean.

3.3 How to Cite

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

3.4 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.

4 Known Bugs

4.1 As of 08/09

- “*Internal Server Error...*” This may occur for a number of reasons. Most likely you have managed to generate a ‘divide by zero’ error. A common situation where this occurs is if you have a population that has missing genotypes for an entire locus. 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. Generally missing data and bootstrapping don’t get along – if you have a lot of missing data or very small population sizes the chance of randomly generating populations that consist of entirely missing data increases.
- ‘*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.
- If you submit a popgen format file with a single locus you can’t have any spaces in the locus name.
- The popgen format expects a header line.
- The server on which I host SMOGD has a max CPU processing time of 2 minutes. So if your data-set takes longer than 2 minutes to analyze, you’ll get an ‘internal server error’. However, I’ve successfully ran data sets with 600 individuals, 10 loci, and 20 populations so this should not be a common problem.

5 Works Cited

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.

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

Jost L. 2008. GST and its relatives do not measure differentiation. *Molecular Ecology* 17(18), 4015-4026. [link]

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(3), 253-259.

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