

TETRASAT

A program for the population analysis of allotetraploid microsatellite data.

Authors

Scott H. Markwith*

David J. Stewart**

and, Jamie L. Dyer†

* Department of Geography, University of Georgia, Athens, Georgia 30601, USA

** Carl Vinson Institute of Government, ITOS Division, University of Georgia, Athens, Georgia 30601, USA

† Department of Geography, Mississippi State University, Mississippi State, Mississippi 39762, USA

INTRODUCTION:

TETRASAT (Markwith et al., 2006) was written for the purpose of calculating genetic diversity, H_E , and H' , and population differentiation, G_{ST} , statistics for populations of allotetraploid species that have been analyzed using microsatellite markers. Due to the exponential amplification process, allele copy number can not be determined for partial heterozygote individuals, thus existing methods of calculating these statistics are largely inadequate. TETRASAT uses an iterative substitution process to determine all possible combinations of genotypes and allele frequencies for each population, and subsequently calculates the necessary statistics. More information concerning the impetus behind the development of TETRASAT, the logic behind its operation, and its utility can be found in Markwith et al. (2006).

INPUT FORMAT:

The input format is similar to that used in the program GENEPOP (Raymond M and Rousset F, 1995). The program is designed to accept space delimited text files, or .prn files, but will accept any text file with the proper spacing. The first line is a file identifier line that can contain any text. The following lines should contain the names of each locus included in the file on separate consecutive lines. Following the locus names should be the text 'Pop' (program is case sensitive on 'Pop'), which is the identifier used to separate each population. In each population each individual should be placed on separate consecutive lines, with the name of the individual or population first, which should include 20 total spaces including text and blank spaces after the name, followed by the tetraploid genotype of the first locus next, followed by a single space, followed by the next locus genotype, followed by a single space, and so on. Since all loci are assumed to be tetraploid, full homozygotes and full heterozygotes individuals should be coded with four two digit alleles per locus, for a total of 8 spaces. Partial heterozygotes should be coded with only the two or three known alleles, and the unknown allele positions left blank, thus there should be 4 or 6 spaces containing the two digit alleles, and 4 or 2 blank spaces immediately following (See Note 1 about the number of partial heterozygotes per population and exceeding computer memory capacity). Individuals completely missing data for a locus should have 8 blank spaces (See Note 2 about how TETRASAT handles missing data). At the end of the input file there should only be one blank line after the genotype data for the last individual. This final line should not have any text or any spaces.

Download and view a sample input file from <http://markwith.freehomepage.com/tetrasat.html>.

RUNNING THE PROGRAM:

1. Download the TETRASAT.exe file from <http://markwith.freehomepage.com/tetrasat.html>. Save to any location on your computer or network.
2. Double click on the TETRASAT.exe file, or right click and select 'open'. The GUI interface will appear.
3. Select the open file icon for 'Formatted input data', and locate your space delimited input text file. The input file can be located in any folder on the computer or network.
4. Designate the file name, file type (.txt, .dat, or .prn), and location to save the 'Output Frequency file', 'Diversity Index Output file', 'Diversity Statistics Output file', and 'G_{ST} Statistics Output file'.
5. Click on the options pull-down menu and select the appropriate options (see below for explanation of options).
6. Press the 'Calc Frequencies' button to start processing the input data. If the 'Automatically execute the next processing step without waiting for input' option has been selected, TETRASAT will complete all possible processes. However, if this option has not been selected, each subsequent process button, 'Calc Diversity', 'Calc Diversity Stats', and 'Calc G_{ST} Values', will need to be pressed by the user upon completion of the previous process.

OUTPUT FILES:

1. Output Frequency file: This is the file where all of the necessary allele substitutions are created and stored as 'allele configurations' for each locus in each population.
2. Diversity Index file: This file includes the allele frequencies, H_E, and H' values calculated from each allele configuration for each locus in each population found in the Output Frequency file.
3. Diversity Statistics Output file: Includes means and standard deviations for H_E, and H' for each locus in each population, and the multi-locus population values.
4. G_{ST} Statistics Output file: Includes means and standard deviations for G_{ST} for each locus in each pair-wise comparison, and the multi-locus pair-wise values.

OPTIONS:

The functions found in the 'Options' drop-down menu are provided to improve the performance of the program. The first option in the list is for opening the 'Options Menu', there are multiple functions in this menu that are explained below. The second option is 'Load from file', this option allows you to load a saved .xml or .xsl file that stores the names and locations of input and output files and chosen options for a previous run of the program. Once the file is chosen, TETRASAT will fill in the appropriate information in the input data, output file, and options boxes. Using this option allows you to duplicate a previous run. The third option is 'Save to file', this option saves the input names and locations of input and output files and chosen options for current run of the program. This option can be chosen prior to or after a run.

Following are options found in the 'Options Menu':

1. Resampling: All resampling strategies are intended to improve the speed of processing and reduce the amount of memory required to process large datasets. If the input data has any one or a combination of the following: 1) more than 10 populations, 2) more than 10 individuals per population, and/or 3) more than 5 loci, then one of the resampling options will likely be necessary to avoid exceeding available memory. Users may want to run the program once without resampling to determine whether memory capacity will be exceeded. (Note: in extreme cases

resampling may be imposed automatically by the application to prevent internal integer overruns; see explanation for option c below about addressing this problem.)

- a. There are two sub-options under the heading ‘When calculating diversity statistics:’. Both of these sub-options only apply to calculation of H_E and H' .
 - i. The first sub-option under this heading is ‘Randomly choose allele configurations for locus-level stats’. A value given for this sub-option will limit the number of allele configurations used to calculate means and standard deviations for any given locus. If a locus has fewer allele configurations than the maximum value entered, then all available configurations will be used at that locus.
 - ii. The second sub-option under this heading is ‘Randomly choose locus-level values to calculate population-level stats’. Population-level multi-locus statistics are calculated taking the mean and standard deviation of a set of all possible multi-locus values created from all possible combinations of locus-level values. When loci each have a large number of values the size of the set of possible multi-locus values can exceed the memory of the computer. Assigning a maximum value for this sub-option limits the total number of multi-locus values that will be created.
- b. There are three sub-options under the heading ‘When calculating G_{ST} values:’. These three sub-options only apply to the calculation of G_{ST} .
 - i. The first sub-option under this heading is ‘Read a random subset of allele configurations from diversity output file’. If there are a large number of allele configurations in the Diversity Index file, enough that the computer runs out of memory simply reading the file, you can randomly choose a sub-set to read. The value entered is the maximum number of configurations per locus. That sub-set forms the pool on which the locus and population G_{ST} calculations are made. The program sets up an array of true/false values, one for each configuration it expects to see in the Diversity Index file. It randomly sets the entered maximum number of those values to TRUE. Then it opens the Diversity Index file, reads each line, throws away the data for FALSE entries, saves the data for TRUE entries, and ends up with an in-memory array of the size of the entered maximum number of configurations. If there is no resampling at this step, all the values are set TRUE. The user would normally use this sub-option and the next independently.
 - ii. The second sub-option under this heading is ‘Randomly choose from in-memory allele configurations for locus-level values’. This option applies to the calculation of G_{ST} values at only the individual locus-level, by limiting the number of pair-wise comparisons of configurations to process for each locus. For example, if there are 725 configurations for locus 3 of pop1 and 27 for locus 3 of pop2, there are potentially $725 \times 27 = 19,575$ total combinations to try. If this sub-option is given a value of 2,000, the application would just pick 2,000 of the possible 19,575 combinations. For this option the program resamples by running through a loop the number times entered by the user. Each pass through the loop it generates a random number between 1 and the number of configurations available in memory. It pulls that data corresponding to that random number out of memory, processes it, and goes on to the next iteration. If there are less pair-wise configuration comparisons in a locus than the entered number, the program processes all possible comparisons and stops.
 - iii. The third sub-option under this heading is ‘Randomly choose locus-level values to calculate population-level values’. The resampling process essentially works the same way as sub-option ii under this heading, however, instead of pulling

data from one locus it pulls data from all loci to create the multi-locus pair-wise population comparisons. For example, if there are 9, 27, and 1 locus configurations at three loci for pop1 ($9 \times 27 \times 1 = 243$), and 9, 1, and 3 locus configurations at three loci for pop 2 ($9 \times 1 \times 3 = 27$), the total number of configurations to try is $243 \times 27 = 6561$. For this option the application would randomly draw from locus pools 1, 2, and 3 of pop1, randomly draw from locus pools 1, 2, and 3, of pop2, calculate a G_{ST} value, and repeat as many times as necessary until the specified limit is reached.

- c. The option 'When an integer overflow is encountered and no sampling limit has been otherwise set, limit arrays to ___ elements' is a random resampling option that has a set default of 10,000 elements. This default can be changed by the user. An integer overflow is a number that exceeds Visual Basics 32-bit number range of -2147483648 to 2147483647. This range is occasionally exceeded during calculation of population-level diversity and G_{ST} statistics for populations or pair-wise comparisons obtained from multiple loci with a large number of partial heterozygotes. This option is a failsafe device that is always on, otherwise the program will crash during an integer overflow.
2. 'Do not display the results panel'. When selected the results panel will not be displayed.
3. 'Automatically execute the next processing step without waiting for input'. When selected all four processes will run sequentially without stopping. When the 'Calc Frequencies' button is pressed, processing time bar and 'Abort' button will appear at the bottom of the TETRASAT GUI display. Pressing the 'Abort' button will stop the program.
4. 'Split diversity index files by population and locus'. When selected the Diversity Index file will be split up into separate files for each locus in each population. This file splitting will occur during the 'Calc Diversity' process if selected, and will help reduce processing time and the possibility of exceeding available memory during processing of the 'Calc Diversity Stats' and 'Calc G_{ST} Values' processes. This option can also be performed outside of the normal processing of input data during the 'Calc Diversity' process. Under the 'File' drop-down menu there is an option called 'Break up output'. When chosen a new window will appear where the name of an existing output file can be specified. This selected file will be split up into separate files for each locus in each population. When the 'Calc Diversity Stats' and 'Calc G_{ST} Values' processes are run individually after this 'Break up output' option is performed, they will recognize the split files and process from those files.
5. 'Display loci and populations by name rather than by number'. When selected the output files will give locus and population values the names of loci and populations as provided in the input file. If not selected, sequential locus and population numbers are given.

NOTES:

1. Using a Dell Optiplex GX280 PC with a Pentium 4 3.00 GHz processor and 1 GB of RAM, TETRASAT could only handle loci that contained 15 or less partial heterozygote individuals. The overall size of the population is not a problem, but it is the number of partial heterozygotes that determines the number of allele combinations. With 15 partial heterozygotes, the number of combinations is 14,348,907. With 16 heterozygote individuals the number of combinations was 43,046,721, and memory capacity was exceeded and the program failed.
2. TETRASAT does account for individuals that have completely missing allele data for one or more loci. As mentioned in the Input Format section, 8 blank spaces should be left at the locus for individuals for whom there is no information. When the program is conducting the iterative substitutions, those individuals with completely missing data at a locus are effectively excluded from the process. For example, at a locus for a population containing 10 individuals with the tenth individual missing information, all possible substitutions will be conducted where necessary

for the partial heterozygotes in the first 9 individuals. To obtain every combination of allele frequencies at that locus, since there are 4 alleles per tetraploid locus, the count of each allele in each combination will be divided by 36 instead of 40. For any other locus associated with that population that the tenth individual does have genotype information that individual are included in the substitutions and calculations.

3. During the review process for the first research paper using TETRASAT (Markwith and Scanlon, 2007), one of the reviewers asked whether the estimates provided by TETRASAT were unbiased. The following is the response provided by Scott Markwith to the reviewer's question:

“For Markwith et al. (2006), the paper introducing the TETRASAT program, allozyme data was used for testing. With allozymes band intensity can be used to determine allele copy number, so the gene frequencies are known for sure. The testing compared statistics that were hand calculated from the complete allozyme dataset, to TETRASAT derived statistics that were calculated from the same allozyme data that was modified to remove certainty of allele copy number. This testing appears to indicate that statistics calculated for data with uncertain band number using TETRASAT are on average somewhat greater than those calculated for data with known band number and certain gene frequencies (The program was also tested using the allozyme data with known allele copy number. Those statistics are identical to the statistics calculated manually, so the statistics are calculated correctly).

The cause of the greater value of the diversity statistics is most likely due to the nature of the statistics themselves combined with TETRASAT's method of averaging all possible combinations. Both expected heterozygosity and the Shannon Index are sensitive to allele frequency evenness. Thus, a tetraploid locus with two copies of allele A and two copies of allele B, AABB ($H_E = 0.50$), will have a diversity value greater than a locus with three copies of allele A and one copy of allele B, AAAB ($H_E = 0.375$) (or one copy of A and three copies of B, ABBB ($H_E = 0.375$)). TETRASAT essentially iteratively creates all three of these genotypes and calculates diversity statistics from each, then takes the average and standard deviation. So the statistic is greater than a genotype with AAAB or ABBB, but less than AABB ($H_E = 0.417$). This is a very simplified example, but I think this is where the bias is generated.

A quick and dirty correlation analysis between deviations of the TETRASAT estimated population parameters from the actual parameters and the number of partial heterozygotes in each of the 15 test populations with two known alleles shows $r = 0.48$ and 0.42 for H_E and H' , respectively. This moderate correlation is not conclusive, but provides some evidence for the hypothesis. The test data is limited, and some of the apparent bias may be due to the small sample size.

In the test data the average bias across all 15 populations is 0.0195 for H_E (range = $-0.0038 - 0.0567$), and 0.0322 for H' (range = $-0.0054 - 0.1105$). The bias for G_{ST} values was lower on average than those for H_E . These values are not zero, but are very low on average. Normally, a researcher would not make a significant conclusion based on a 0.0195 difference in expected heterozygosity. The inclusion of standard deviations with every mean expected heterozygosity value reported is intended to give an indication of variability around the mean. Because the variability and the bias are both created in the same process the bias is actually a part of the standard deviation. In addition, because all TETRASAT derived statistics are subject to the same processes, and thus the same systematic bias, comparison of parameters that are all calculated using TETRASAT should not be problematic.”

CORRECTIONS TO M.E.N. TETRASAT ARTICLE:

The Markwith et al. (2006) *Molecular Ecology Notes* article concerning TETRASAT has a couple of mistakes, the corrections follow:

1. The equation for the Shannon-Weiner Diversity Index is missing a negative sign. The correct equation is:

$$H' = -\sum p_i^2 \log_2 p_i$$

2. The references are missing a citation. The missing citation is:

Nei, M. 1986. Definition and estimation of fixation indices. *Evolution*, vol. 40, p. 643-645.

REFERENCES:

Markwith, S. H. and Scanlon, M. J. (2007). Multi-scale Analysis of *Hymenocallis coronaria* Genetic Diversity, Genetic Structure, and Gene Movement Under the Influence of Unidirectional Stream Flow. *American Journal of Botany*, vol. 94, no. 2, p. 151-160.

Markwith, S.H., Stewart, D.J., and Dyer, J.L. (2006). TETRASAT: A program for the population analysis of allotetraploid microsatellite data. *Molecular Ecology Notes*, vol. 6, p. 586-589.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, 86, 248-254