

# **GenoType/GenoDive**

Applications for analysis of genetic diversity of asexual organisms

User's manual

31-aug-2006

Patrick Meirmans  
I.B.E.D.  
Universiteit van Amsterdam

Kruislaan 318,  
1098 SM, Amsterdam,  
The Netherlands

<http://www.science.uva.nl/~meirmans>

## GenoType and GenoDive

GENOTYPE and GENODIVE are two relatively small applications for analyzing genotypic diversity in populations of asexually reproducing organisms, using molecular data.

These applications are meant to be the forerunners of a future, more complete, software program for analyzing asexual organisms. That program will among others also be able to do a Character Incompatibility Analysis (Mes, 1998), tests for linkage disequilibrium and more elaborate analyses of detecting "hidden" sex in asexual organisms.

If you have used one of the programs, or both, please cite the following reference:

Meirmans P.G., Van Tienderen P.H. (2004) GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, 4, 792–794

If you have any comments or problems please do not hesitate to contact me at: [meirmans@science.uva.nl](mailto:meirmans@science.uva.nl).

Good luck! Patrick Meirmans

## GenoType

GENOTYPE is a program for assigning genotypic identity to individuals, using data from most types of genetic markers. Identification of genotypes is especially important in studies of parthenogenetically reproducing organisms, or with organisms with clonal reproduction. In these cases different individuals (sometimes referred to as "ramets") can have the identical multilocus genotypes (referred to as "genets"). In studies with noninvasive sampling, GENOTYPE can also be used to check whether some individuals have been sampled twice. Assigning genotypes is normally a relatively simple task, but for large datasets it is tedious to do it by hand. There are other programs that can do perform similar tasks (e.g. Fstat, Goudet, (1995), Gimlet, Valière (Valière, 2002)), but only for diploid data; Our program has specifically been designed to handle polyploid data, as a large part of asexual organisms are polyploid. Furthermore, GENOTYPE takes into account and controls for scoring errors and mutations, which can cause different ramets from the same genet to be slightly different (Douhovnikoff, Dodd, 2003).

## GenoDive

GENODIVE is a program for calculating and testing indices of clonal diversity, meant to be used after assigning genotypes to individuals with GENOTYPE. GENODIVE calculates different diversity indices such as the number of genotypes, the effective number of genotypes, Nei's (1987) diversity index (identical Simpson's diversity index), the corresponding evenness, and the, corrected, Shannon-Wiener diversity index . It can also perform a bootstrap test to see whether these indices are different for pairs of populations. Furthermore, the program can perform jackknives with increasing sample sizes to check whether the population sample sizes are big enough to have an unbiased estimate of clonal diversity. Finally, the program can also test for differentiation in genotypic composition between pairs of populations.

## Input files

### GenoType:

The input file of GENOTYPE should be a tab delimited text file, with the specifications detailed below. There are no limitations to the numbers of individuals, populations or loci, but there is a maximum ploidy level of 16 (hexadeciploid). Note that the speed of calculation is mainly dependent on the size of the distance matrix, which can be very big for large datasets. GENOTYPE can also read input files in Fstat format (Goudet, 1995): files with the extension *.dat* are assumed to be in Fstat format, all other files are assumed to be in GENOTYPE format. Unlike the Fstat-program itself, GENOTYPE can also read files in (a somewhat extended) Fstat format containing polyploids.

The GENOTYPE format is as follows:

-First line: comments (obligatory, maximum of 200 characters).

-Second line (separated by tabs):

number of individuals  
 number of populations  
 number of loci  
 maximum ploidy level present within the data set  
 number of digits used for coding an allele

-Names of populations (should match the number of populations, separated by returns, maximum of 25 characters per name, no spaces are allowed)

-Per individual (separated by tabs):

population number (from 1 to n)  
 name of individual (max 10 characters, no spaces)  
 alleles (per locus as one string)

If an allele name starts with a zero, it is possible to skip the first zero for a locus (e.g. genotype 0102 can be given as 102). Missing data should be entered as follows: If all alleles for a certain locus are missing, a zero should be entered. If only some alleles are missing (e.g. due to null-alleles, PCR-artifacts or scoring problems), all known alleles should be entered; the program recognizes the missing alleles due to the apparent difference in ploidy level compared to the other loci of the same individual. After the allelic data from the last individual, it is possible to add further comments.

A problem I encountered when making input files with Excel is that Excel sometimes adds several tab characters at the end of a line. These tabs may cause population names to be longer than allowed and cause the program to crash. This seems to be the case only when a huge amount of loci is used. In such cases, try to remove unnecessary tabs with a text editor.

An example input file, called `example_genotype_msat.txt`, with microsatellite data can be found in the same folder as the GENOTYPE and GENODIVE programs.

```

example input file (includes this comment line)
4      2      2      3      2
pop1
pop2
1      John      102      1214
2      Paul      202      0
1      George    101      121213
2      Ringo     10304    131414

```

### ***Inputting data from dominant markers***

Binary scored (dominant) data, such as RAPD's or AFLP's, should be entered as if the species were haploid, with a separate locus for every marker, and with one-digit alleles. As zeros are reserved for missing data (described above), absence/presence has to be entered in a different way than the standard 0/1. Use, for example, an 8 for absence and a 9 for presence, or 7 and 3, or 2 and 6, or any other combination.

```

example input file (dominant data)
4 2 5 1 1
pop1
pop2
1 John 8 8 9 9 8
2 Paul 9 8 9 9 8
1 George 9 0 9 9 9
2 Ringo 9 8 9 9 9

```

An example input file, called `example_genotype_aflp.txt`, with dominant data can be found in the same folder as the GENOTYPE and GENODIVE programs. This example consists of AFLP-data from a sample of 67 dandelions from Viborg, Denmark, kindly provided by Ron van der Hulst (Van der Hulst *et al.*, 2003).

### ***Importing a distance matrix***

It is possible to import a distance matrix if you want to use another distance index than the ones offered by the program. The imported matrix should be a tab delimited text file with the distances in either a square or unfolded lower triangular format. All distances should be entered as integers, so the case of a distance index that is bounded between 0 and 1, it is best to multiply everything by a hundred. Note that when you use your own distance matrix, the program still requires a normal input file, though all the allelic data will be ignored. Furthermore, the number of individuals in the distance matrix should match those in the standard input file. Here is an example of a square distance matrix:

```

0      20      33      33
20     0       14      14
33     14      0       0
33     14      0       0

```

**GenoDive:**

The input file of GENODIVE should be a tab delimited text file, and should be placed in the same folder as the program. The format of the input file is as follows:

-First line: comments (obligatory, maximum of 200 characters).

-Second line: number of populations.

-Names of populations (should match the number of populations, separated by returns, maximum 20 characters)

-Per individual (separated by tabs):  
population number (from 1 to n)  
genotype number

An example of an input file is given below. An example input file, called `example_genodive.txt` can be found in the same folder as the GENOTYPE and GENODIVE programs.

```
example input file (includes this comment line)
2
pop1
pop2
1      1
2      2
1      1
2      3
```

Like GENOTYPE, GENODIVE has no restrictions regarding the number of populations and individuals it can handle.

## Using GenoType

The program starts by asking for the name of the input file, this file should be placed in the same folder as the program. The full name of the input file must be given including the extension (usually *.txt*). Files with the extension *.dat* are assumed to be in Fstat format. If the data has been read correctly the program gives short summary of the data.

### Distance options

Next, GENOTYPE asks which distance index should be used; the program works through calculating the genetic distance between all pairs of individuals. Different distance indices are available which can be chosen from the following menu:

Calculate distance matrix:

1. Import matrix from file
2. Stepwise mutation model
3. Infinite allele model
4. Dice similarity (as a distance, in percent)

#### *Import matrix from file (distance option 1):*

If you want to use a distance index different from the ones offered by the program, you can import your own distance matrix (see the section about the input format above).

#### *Stepwise mutation model (distance option 2):*

This distance index is specifically meant for microsatellite data. A stepwise mutation model (SMM) is assumed, meaning that alleles that differ only a few repeats in length are thought to be of more recent common ancestry than alleles that differ a lot of repeats in length. For proper calculation, this index requires that the alleles are given as the number of repeats rather than the length of the fragment. However, also fragment length data can be given, which is handy for microsatellite loci containing imperfect repeats, though this stretches the idea of the SMM a bit. The program has different ways to calculate the distances, you can choose the method from the following menu:

Number of mutation steps

1. Missing data not counted
2. Missing data equals specified number of mutations

Squared number of mutation steps (Amova-type)

3. Missing data not counted
4. Missing data equals specified number of mutations

Number of mutation steps, scaled by ploidy (two-phase model)

5. Missing data not counted
6. Missing data equals infinitely large mutation
7. Specify settings

The most straightforward distance method under the SMM is to simply calculate the smallest number of mutation steps that is needed to transform the genotype of one individual into the genotype of the other, summed over all loci (options 1 & 2). For haploid data without missing values, this distance measure is equivalent to the Manhattan distance. Options 1 and 2 differ in the way that they treat missing data: under option one missing data are discarded, under option two missing data equal a certain number of mutation steps, provided by the user (in that case a low number should be

given, otherwise the maximum possible distance will get very high in datasets with lots of missing data). Under option two it is also possible to let missing data be substituted by the average allele length for a locus.

The second method of calculate distances under the SMM, is for use with AMOVA's. Here the overall distance is the sum of squared distances. For the rest the options are the same as above.

The third method of calculate distances under the SMM, is to scale the distance between 0 and 1, taking the ploidy levels into account (options 5, 6 & 7). This method (Bruvo et al. *in press*) uses the two-phase model (Di Rienzo *et al.*, 1994), which can be seen as a more specific version of the SMM. Under the two-phase model, long mutation steps are less likely to occur than shorter ones; the likelihood of the mutation is exponentially related to the step length. Note that the program determines the ploidy level of an individual from the maximum number of alleles the individual has over all loci. So for proper calculation, every individual should have no missing data for at least one locus. To make distances better viewable in the histogram, all distances are multiplied by a hundred and rounded to integers. With this method missing data can be handled in different ways: they are either discarded (option 5), equated to an infinitely large mutation step (option 6, which one of the three methods described by Bruvo et al.), or equated to a mutation step of a different length, specified by the user (option 7). With the last option, it is also possible to specify the base of the exponent, the default of which is two (as in  $2^n$ , where  $n$  is the number of mutation steps). With option 7 it is also possible to specify by what number the distances are multiplied before they are rounded (the default is 100), this is handy if you are not interested in distinguishing genotypes, but in calculating the distances. To this end you should use a large multiplication factor (e.g. 10.000) in order to avoid rounding errors.

*Infinite allele model (distance option 3):*

Under this distance index, an infinite allele model (IAM) is assumed meaning that it takes only one mutation step to get from a certain allelic state to any other. This mutation model is valid for almost all molecular data besides microsatellites, such as allozymes, RAPD's, and AFLP's. The distance measure simply consists of the number of mutations that are needed to transform the genotype of one individual into the genotype of the other, summed over all loci. Like the simple SMM-distance, the IAM distance is equivalent to the Manhattan distance for haploid data without missing values.

Infinite allele model, missing data:

1. not counted
2. counts as one mutation
3. counts as specified number of mutations

The program can handle missing data in three different ways: they are either discarded (option 1), equated to one mutation (option 2) or equated to a user-specified number of mutations (option 3).

*Dice similarity (distance option 4):*

This option is only available for dominant data; the program only shows the option when the input consists of binary coded, haploid, data. The Dice similarity index is calculated between two individuals as  $2a/(2a+b+c)$ , where  $a$  is the number of bands

shared between the individuals, b is the number of bands present in the first individual but not in the second, and c is the number of bands present in the second individual but not in the first. The Dice similarity is normally bound between 0 and 1; when it is 0 the individuals are completely different, when it is 1 they are identical. GENOTYPE transforms this similarity index into a distance by subtracting it from one, then it multiplies it by hundred for better representation in the histogram. Before calculation of the distance matrix, an additional question must be answered about which number should be interpreted as coding for 'absence' (see also the above part about inputting dominant data; in the example file '8' is used as the number coding for absence).

## Histogram

GENOTYPE prints a very basic histogram: The frequency distribution of all the pairwise distances. An example of such a histogram is given below:

class	thresh	#types	het_w	#pairs	#pairs
0	0.00	42	0.00	62	***
1	3.00	31	0.46	41	**
2	6.00	28	0.52	6	
3	9.00	21	0.61	71	****
4	12.00	7	0.53	103	*****
5	15.00	2	0.98	338	*****
6	18.00	1	0.97	454	*****
7	21.00	1	0.97	596	*****
8	24.00	1	0.97	355	*****
9	27.00	1	0.97	117	*****
10	30.00	1	0.97	57	***
11	33.00	1	0.97	9	

The program asks for the number of classes that the histogram should contain, and the maximum distance in the histogram. Usually these two should be the same and equal to the maximum distance in the dataset, which is given right before the program asks this question. However, an incomplete histogram can be useful if you are not interested in the full range of distances, but rather want to focus on short distances: the number of mutations between closely related individuals.

The "threshold" concept is of vital importance in the program. The threshold indicates the maximum distance that is allowed between two individuals to still be clonemates with the "same" multilocus genotype. Scoring errors and mutations may cause individuals from the same clonal lineage (clonemates) to have a pairwise distance larger than zero. To set a limit to this you can draw a threshold for the amount of scoring error or mutation you allow. Choosing this threshold too low inflates the estimates of clonal diversity, choosing this value too high deflates the diversity estimates, so choosing a right threshold is important. Douhovnikoff and Dodd (2003) recently proposed a method to objectively choose a threshold value, based on the means and standard deviations of the two peaks in a bimodal histogram. This method is however not implemented in GENOTYPE as it is difficult to implement on data from natural populations. Douhovnikoff and Dodd however assumed the determination of the threshold from a dataset of known clones and siblings, and afterwards used this value for natural populations. As most studies on clonal diversity will probably be carried out without the possibility of such prior testing, I did not include this method in the program, but rather propose to test hypotheses concerning clonal diversity using several different threshold-values, to see the effect of the scoring errors and mutations on the used statistics.

From left to right are shown in the histogram: the class number, the corresponding



threshold, the number of genotypes that are distinguished using that threshold, the average diversity within the genotypes and the number of pairwise distances that are found within the class. The latter is depicted both numerically and graphically; note that the number of genotypes reaches one at a certain threshold. The diversity index can be seen as an indicator for the number of different multilocus genotypes (those which would be recognised under threshold = 0) that are grouped under one genotype for the current threshold. The calculations for the diversity, and also for the number of genotypes, can be slow for large datasets. Therefore, for datasets containing more than 500 individuals, the program asks whether you want to speed up the drawing of the histogram; in that case the program does not calculate the number of genotypes and the diversity.

## **AMOVA**

In version 1.2 GENOTYPE can perform a very simple AMOVA (Excoffier et al. 1992), which calculates the amount of differentiation among populations, without any higher-level grouping or within individual level. The AMOVA uses the distance matrix that was calculated in the previous step. For a correct inference of  $\Phi_{st}$ , the distance matrix should contain squared Euclidean distances (so use any of the IAM distances or the squared SMM distance-option). The AMOVA also performs a permutation test by randomising individuals over populations; simply give the number of permutations when asked for it. For big datasets this test can take a while and there is no progress indication, so be patient.

## **Output**

Following the drawing of the histogram, the program asks whether to try it again, meaning recalculating the distances and the histogram, or whether to proceed to the output-menu. GENOTYPE can save the output in five different formats: an input file for the GENODIVE program for one particular threshold, a file with genotypes for one particular threshold, a file with genotypes for all thresholds, a file with the distance matrix and a file with the histogram. If you specify a name for an outfile, the program will automatically overwrite any other file with the same name in that directory, so choose a unique name unless you want the file to be overwritten.

### *GENODIVE input file:*

This output option saves a file in the format that is used as an input file by the GENODIVE program. You will be asked a threshold-value, which is used to assign genotypes to individuals, and you will be asked to give a name for the outfile.

### *Genotypes file:*

This is simply a file with for every individual its name and the genotype, according to a certain, user-specified, threshold-value.

### *Genotypes for all thresholds:*

This outfile makes the previous one completely redundant. It gives for every individual the genotype it belongs to under all relevant threshold-values (those that give more than one genotype in total).

### *Distance matrix:*

This output option writes the distance matrix that is used for assigning the genotypes to a file. You can choose between two different formats: a square matrix and an unfolded

lower triangular matrix (without the diagonal). The latter option may be useful for large datasets as, for example, Microsoft Excel is not able to open square matrices with more than 256 columns (although it also can not open unfolded matrices calculated from more than 362 individuals).

*Histogram:*

The output file contains a histogram, which is produced in the same way as described above, though the histogram is written to a file rather than to the screen.

## Using GenoDive

When starting up GENODIVE, the program asks for the name of the input file. The input file should be placed in the same folder as the program and should be a tab-delimited text file. You must type the full name of the input file including the extension (usually *.txt*). If the data has been read correctly the program provides a short summary of the data.

Next, you are asked for the name of the output file, which will be stored in the same folder as the program. If a file with the provided name already exists in that directory, the program will append the output at the end of the file.

## Analyses

GENODIVE can perform several different types of analyses, which can be chosen from the following menu:

```

Statistics
  1: Calculate indices of clonal diversity (default)
  2: Bootstrap test for differences in clonal diversity
  3: Pairwise test for population differentiation
  4: Jackknifing with increasing sample sizes

Options
  6: Write input-file in Fstat-format
  7: Open new input and output files
  8: Set seed for random number generator
  9: Exit program

```

### *Calculate indices of clonal diversity (option 1)*

When this option is chosen, the program calculates for every population the following indices of genotypic/clonal diversity. In the formula's  $s$  is the number of genotypes,  $n$  is the sample size,  $p_i$  is the frequency of genotype  $i$  in the sample and  $u$  is the number of genotypes that are only present once. The three digit abbreviations of the diversity indices are used in the output and interface of GENODIVE.

1. The number of genotypes (num).  
Simply the number of genotypes found in a population ( $s$ ).

2. The effective number of genotypes (eff):

$$\frac{1}{\sum_{i=1}^s p_i^2}$$

This is equivalent to the "effective number of alleles" that is sometimes used in allozyme studies. This index may be slightly more insightful than the number of genotypes for comparing diversity between populations, though care should be taken as this index is biased for small sample sizes.

3. Nei's (1987) genetic diversity (div) corrected for sample size:

$$\frac{n}{n-1} \cdot \left(1 - \sum_{i=1}^s p_i^2\right)$$

Among ecologists, this index is better known as Simpson's diversity index. This is the only diversity index calculated by GENODIVE that is truly independent of sample size. The index is also widely used in population genetics under the name of "expected heterozygosity".

4. The evenness (eve):

$$\frac{1}{S} \cdot \frac{1}{\sum_{i=1}^s p_i^2} \quad (= \text{eff} / \text{num})$$

Basically, every diversity index has its own evenness, which is simply calculated as the estimated value of the index divided by the maximum value possible for the used sample size. For a diversity index that has a maximum of one, the corresponding evenness therefore equals the index itself. The evenness is an indicator for how evenly the genotypes are divided over the population, hence the name. An evenness value of 1 indicates that all genotypes have equal frequencies. GENODIVE calculates the evenness of the effective number of genotypes, which is the most widely used one. However, as the effective number of genotypes itself it has an estimation bias related to the sample size, also the related evenness has an estimation bias.

5. Shannon-Wiener index (shw):

$$-\sum_{i=1}^s p_i \cdot \log p_i$$

This index, also known as the Shannon-index or as the Shannon-Weaver index, is the most widely used diversity index in ecology. However, the index has a huge estimation bias and therefore is not always useful in genetics, unless all sample sizes in a study are more or less equal. The corresponding evenness can easily be calculated by dividing the estimate by  $\log(s)$ . For calculating the Shannon-Wiener index, some people use a natural logarithm instead of  $^{10}\log$  that is used here, so be careful when comparing your results with other studies unless you know what method they used.

6. Shannon-Wiener index (shc) corrected for sample size (Chao, Shen, 2003):

$$-\frac{\sum_{i=1}^s C \cdot p_i \cdot \log(C \cdot p_i)}{1 - (1 - C \cdot p_i)^n}, \text{ where } C = 1 - u/n$$

This is a recently published version of the Shannon-Wiener index that uses a non-parametric bias correction. For the correction, the number of singletons (types that are only sampled once) is used to estimate the number of unsampled types. Though this removes the bias rather well for sample sizes  $> \sim 50$ , it still has a bias for smaller sample sizes. The used method of correcting the bias is not possible in some cases: e.g. when all the individuals in a populations have different genotypes. In that case, you will see "nan", instead of a number.

7. Nei's uncorrected genetic diversity (diu):

$$1 - \sum_{i=1}^s p_i^2$$

This is index number 3, without the  $n/(n-1)$  correction term.

Next to these population specific indices, for Nei's index and the two Shannon-Wiener indices estimates of the total diversity, the average diversity per population and the among-populations diversity are given. However, there is a difference in calculation of these estimates between the different indices. For Nei's index, I used the formulas from Nei (1987), and the fraction of among-populations diversity is in this case equivalent to  $G_{st}$ . For the two Shannon-Wiener indices, I found no ways of properly calculating the total and average within population diversity. The total diversity is therefore simply calculated by pooling all individuals and the average within-population diversity is calculated through averaging over populations. This method is not free of bias: The Shannon-Wiener indices depend on the sample size and the total sample size is always bigger than the average of the population sample sizes. Therefore, the estimate of the total diversity will usually be higher than the estimate of the within-population diversity, even when there should be no among-populations diversity (e.g. when all the samples come from the same population). In other words: the estimate of the fraction of among-population diversity gets inflated. This bias is, however, less for the corrected Shannon-Wiener index, as it is less dependent on the sample size.

For convenience, the program shows the among-population components of the Shannon-Wiener indices also under " $G_{st}$ " in the output, though these are technically no real  $G_{st}$ -values (and certainly not estimates of  $F_{st}$ !). These " $G_{st}$ " estimates are not corrected for the number of populations, and therefore not well suited for datasets containing only a few populations. Therefore, also the corrected versions " $G'_{st}$ " are calculated (see Nei 1987). Note that all the " $G_{st}$ " and " $G'_{st}$ " values are based on genotype frequencies and not on allele frequencies!

The " $G_{st}$ " and " $G'_{st}$ " estimates for Nei's index are usually lower than those for the Shannon-Wiener indices. This is not only because of the above-mentioned bias, but also (mainly) because of the different (statistical) behaviour of the indices. This difference in behaviour is the reason that most ecologists prefer the Shannon-Wiener index, even though it has an estimation bias; they think that the Shannon-Wiener better represents the diversity seen in nature. It must be said, however, that in Ecology the estimation bias is generally not as important as it is in Genetics.

#### *Bootstrap test for differences in clonal diversity (option 2)*

This method tests whether pairs of populations differ in their clonal diversity. This type of test can be especially useful if some populations are expected to have a higher diversity than others; for example due to the presence of sexuals in these populations or because of an expected geographical trend in clonal diversity. The test uses a bootstrapping approach (resampling with replacement); the individuals are resampled from the populations and the diversity indices are compared after every replicate (Manly, 1991). GENODIVE asks for the number of permutations (for example 1000 or 10.000) and whether to save all the permuted values of the statistics into a file; this enables you to check their distribution. The test is performed for all diversity indices (but not pairwise population comparisons!) simultaneously, which makes the p-values for the different diversity indices dependent on each other. However, I assume that most users will focus on only one index and ignore the others. The bootstrap test has a bias when the sample sizes of the compared populations differ a lot and of course the test also has a bias for diversity indices with an estimation bias. These biases can be overcome by subsampling the samples to be of equal size before creating an inputfile. There recently has been a discussion on Evoldir about whether it is better to use the bootstrapping technique or Nei's (1987) method by using the variance of his diversity

estimate. A brief comment on this is also given in the methods-section in Thomas et al. (Thomas *et al.*, 2002). In my view, testing hypotheses on trends in clonal diversity preferably requires tests between groups of populations rather than testing between pairs of populations.

#### *Pairwise test for population differentiation (option 3)*

This method tests whether pairs of populations differ in their genotypic composition, or rather: whether genotype frequencies differ between two populations. This is similar to testing for differences in allele frequencies, which is possible in most population genetic programs, only in this case the alleles have been replaced by genotypes. This method should not be confused with the bootstrap test above, which tests for differences in genotypic *diversity* rather than genotypic *composition*. The test statistic used is the log-likelihood ratio G-statistic, the significance of which is determined by randomizing genotypes among populations. When performing this test you can specify the number of permutations (for example 1000 or 10.000) and whether or not you all the permuted values of the statistics should be saved into a file.

#### *Jackknifing with increasing sample sizes (option 4)*

This method can be used to check whether sample sizes are big enough to be able to estimate some of the diversity indices without bias. Some of the indices used (even the "corrected" Shannon-Wiener index) have an estimation bias for small sample sizes. With the jackknife method you can check whether your sample sizes are big enough to avoid such bias by taking increasingly large subsamples of your data, starting at 2. If the trend in the value of the index for the different subsample sizes has leveled off when it reaches the actual sample size, the sample size was adequate. If the trend has not leveled off, you should have sampled more individuals, or you should calculate an unbiased diversity index (Though arguably you may need to calculate a biased index for comparison with other studies that used that index, but that comparison may be nonsense simply because of the bias). When performing this test you can specify the number of permutations per jackknifed sample size (for example 1000 or 10.000).

#### *Other options (options 6 to 9)*

These are rather self-explanatory. Option 6 will save the input file as an Fstat (Goudet, 1995) input file, which is useful if you want to do further tests on the genotype frequencies or calculate F-statistics other than the G-statistic calculated by GENODIVE. Note that if you work on a Mac, Fstat will not read the data unless you convert the "line-breaks" to Windows format first. You can do this by opening the file in Excel and save it as Text (Windows). Option 8 will set the seeds for the random number generator; by default, the system clock is used for setting the seeds.

## **Output**

GENODIVE is able to produce a pretty large amount of output even for small datasets, especially if you choose to save the permutations to a file. The exact content of the output file depends on the kinds of tests you chose to do. All p-values in the output file are given without any correction for multiple testing (Bonferroni correction).

#### *Indices of clonal diversity:*

The output file shows for every population the indices of clonal diversity it calculated. Those indices are described above in a bit more detail. If there is more than one

population, the total diversity is partitioned into within and among population components ( $G_{st}$  and  $G'_{st}$ ).

*Bootstrap tests for differences in genotypic diversity between populations:*

The output file shows for every combination of two populations (labeled *pop A* and *pop B*) for six diversity indices whether the two populations differ for these indices. Per index, two p-values are given:  $p(A \geq B)$  and  $p(A \leq B)$ . This allows you to test either one-sided or two sided. In a two sided test, the difference in diversity is significant if one of the two p-values is smaller than 0.025. If you have an *a priori* hypothesis about which population should have the highest diversity, you can test one-sided and it is significant when the p-value corresponding to your hypothesis is smaller than 0.05. You should be careful with drawing conclusions from these bootstrap tests: they have a bias, which leads to an inflated type 1 error. This is because when you are resampling with replacement, as is done when bootstrapping, the diversity in your bootstrapped sample will almost always be lower than in your original sample. This effect is dependent on the sample size: it will be worse with small populations. So the test is most appropriate when the sample sizes of the compared populations are more or less equal. Next to this, the diversity indices that have a bias for sample size (*gen*, *eff*, *eve*) are not really suited for this test. The evenness also has a bias for the number of genotypes in a population (next to a bias due to sample size), so the bootstrap test for this may only be informative when the both the sample sizes and the number of genotypes in the compared populations match.

The file with the permuted values of the bootstrap test-statistic gets very big for a large number of permutations or a large number of populations; sometimes too big for Excel to read it. In the file there are three times six columns with data. The first six columns consist of the permuted values of the diversity indices for population A, the next six columns those of population B and the last six columns consist of the values of the test statistic used for the bootstrap test:  $(\text{permA} - \text{permB}) - (\text{origA} - \text{origB})$ . For every pairwise combination of populations, the first line of the eighteen columns contains the original values, the next (permutations - 1) lines contain the permuted values.

*Test for population differentiation:*

The output file simply shows a pairwise matrix with for every combination of populations the p-value of the permutation test. The file with permuted values gives for every pair of populations on the first line the original G-statistic (note that this is the log-likelihood G-statistic, not  $G_{st}$ ), and on all consecutive lines the permuted values of the G-statistic, that is calculated after randomising the genotypes over the two populations.

*Jackknifing with increasing sample size:*

The outputfile shows for increasing subsample sizes (from 2 to  $(n-1)$ ), the results of the jackknife permutations for the five diversity indices. Per index are given first the average of all permuted values, and then the lower and upper bounds of the 95% interval around this average. This means that 95% of the permuted values lie between these two bounds. For low subsample sizes, the corrected Shannon Wiener index will mostly give the indication *nan* (not a number). This means that for at least one of the permuted datasets the index could not be calculated due to the non-parametric mode of bias-correction (Chao, Shen, 2003). The last row for every population gives the diversity indices for the complete sample, these are the same as those calculated under option 1.

## Troubleshooting

The two programs are probably not free of bugs, despite all our testing. However, most problems will not be due to bugs, but will have different causes (most probably related to an incorrect input file). If you have problems other than those outlined below, contact me at [meirmans@science.uva.nl](mailto:meirmans@science.uva.nl).

### *-The program cannot find the input file*

Make sure that you typed the full name of the input file, including the extension. On some operating systems, it may not be obvious that the filename actually has an extension. If the program cannot find a file with the given name, and the name does not have an extension, the program asks whether to open a file with that name and the extension *.txt*.

### *-The program crashes:*

This probably happened just after entering the name of the input file. Check whether the inputfile is conform the specifications mentioned above. GENOTYPE expects that files with the extension *.dat* are in Fstat format, the program may crash if a *.dat*-file is in GENOTYPE format. If you use different operating systems, make sure the line-breaks in the input file are conform the system you run the program on.

### *-The program freezes:*

If you have a large dataset, the program is probably still calculating. GENOTYPE works with distance matrices; these get very big for datasets containing thousands of individuals, so the time required for the calculations can get long. If the calculation of the histogram gets slow, choose the "speedy" way of calculating (only possible with >500 individuals). GENODIVE can also require a long time for its calculations when you use the "jackknife with increasing sample size" option, or if you do a large number of permutations in the pairwise tests. If your computer has not much RAM-memory, it may help to close other programs before starting up GENOTYPE or GENODIVE, to prevent the computer from slowing down due to using virtual memory.

### *-The program gives an error message*

Both programs have two main types of error messages: those related to reading/writing the input and output files (e.g. ERROR: input file error, individual 10, locus 3) and those related to memory (e.g. ERROR: pop\_names malloc failed, more memory needed). If you get an error of the first type, make sure you have typed the name of the file correctly, that the file is not in use by another program and that the input file is conform the specifications mentioned above.

Memory related errors can also be due to an incorrect input file. If this is the case, the error will show up directly after reading the input file. On Macintosh computers running system 9 or below, memory errors can also show up if too little memory has been allocated to the program. To allocate more memory, select the program in the Finder and choose "Get info" from the File-menu. In the info-panel that appears, increase the amount of memory given in the box labeled "preferred size". Also be sure to close other programs before starting up GENOTYPE or GENODIVE.



## Acknowledgements

The two programs wouldn't have been the same without the discussions I had with Ron van der Hulst, who also kindly provided the AFLP data for one of the example files. Thure Hauser had some helpful suggestions for improving the GENOTYPE program. Peter van Tienderen co-authored the manuscript for the program note written about these programs. Last, but not least, Stephanie Hamm corrected a lot of bad English in this manual.

## References

- Chao A, Shen T (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics* **10**, 429-443.
- Di Rienzo A, Peterson AC, Garza JC, *et al.* (1994) Mutational Processes of Simple-Sequence Repeat Loci in Human- Populations. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 3166-3170.
- Douhovnikoff V, Dodd RS (2003) Intra-clonal variation and a similarity threshold for identification of clones: application to *Salix exigua* using AFLP molecular markers. *Theoretical and Applied Genetics* **106**, 1307-1315.
- 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.
- Goudet J (1995) FSTAT (Version 1.2): A computer program to calculate F- statistics. *Journal of Heredity* **86**, 485-486.
- Manly BFJ (1991) *Randomization and Monte Carlo methods in biology* Chapman & Hall, London.
- Mes THM (1998) Character compatibility of molecular markers to distinguish asexual and sexual reproduction. *Molecular Ecology* **7**, 1719-1727.
- Nei M (1987) *Molecular Evolutionary Genetics* Columbia University Press, New York.
- Thomas MG, Weale ME, Jones AL, *et al.* (2002) Founding mothers of Jewish communities: Geographically separated Jewish groups were independently founded by very few female ancestors. *American Journal of Human Genetics* **70**, 1411-1420.
- Valière N (2002) GIMLET: a computer program for analysing genetic individual identification data. *Molecular Ecology Notes* **2**, 377-379.
- Van der Hulst RGM, Mes THM, Falque M, *et al.* (2003) Genetic structure of a population sample of apomictic dandelions. *Heredity* **90**, 326-335.