AFLP-SURV 1.0

A program for genetic diversity analysis with AFLP (and RAPD) population data

written by Xavier Vekemans

Address for correspondence:
Laboratoire de Génétique et d'Ecologie Végétales
Université Libre de Bruxelles
1850 Chaussée de Wavre
B-1160 Bruxelles Belgium
e-mail: xvekema@ulb.ac.be

Last update: 03/06/02

Contents

1. NOTE ABOUT *AFLP-SURV* 1.0

AFLP-SURV 1.0 is a first test version released and is thus very likely to contains bugs. Some of these bugs are probably easy to detect by causing a program crash or leading to crazy results for particular data sets and analyses. But others, more critical, may just cause biased results that however appear plausible. Hence, it is advised to take much care checking the consistency of the information from the output file and, whenever possible, to verify the results with an alternative software. The author would appreciate to be informed of any detected bug (please send a sample data file). The responsibility of the author is not engaged whenever a bug caused a misinterpretation of the results given by *AFLP-SURV*. Any suggestion for improvement of the scope and/or of the functionality of the program is welcome.

AFLP-SURV replaces an older program (RAPD-survey) that performed the analysis of Lynch & Milligan (1994) on RAPD data.

How to cite AFLP-SURV?

Vekemans, X. 2002. AFLP-SURV version 1.0. Distributed by the author. Laboratoire de Génétique et Ecologie Végétale, Université Libre de Bruxelles, Belgium.

If the journal ask you for a printed publication, you can put instead:
Vekemans X., T. Beauwens, M. Lemaire and I. Roldan-Ruiz, 2002. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Molecular Ecology, 11, 139-151.

2. WHAT IS *AFLP-SURV*?

2.1. Purpose

AFLP-SURV estimates **genetic diversity** and **population genetic structure** from **population samples** analysed with **AFLP** or **RAPD** methods and computes **genetic distance** matrices between populations. The program starts by estimating allelic frequencies at each marker locus in each population assuming they are dominant and have only two alleles (a dominant marker allele coding for the presence of a band at a given position, and a recessive null allele coding for the absence of the band). The user has to specify whether **Hardy-Weinberg** genotypic proportions can be assumed, or in contrast whether the organism is completely **homozygous** at marker loci (highly self-fertilising species or haploid species), or alternatively whether there are some **known deviations from Hardy-Weinberg** genotypic proportions. Based on these estimates of allelic frequencies, the program uses either of two approaches to estimate genetic diversity and population genetic structure:

1) The **approach of Lynch & Milligan** (1994) which uses the average expected heterozygosity of the marker loci, or Nei's gene diversity, as a measure of genetic diversity.

2) The approach of Clark and Lanigan (1993) proposed for RAPD loci and extended to AFLP loci by Innan et al. (1999), which uses the nucleotide diversity estimated from the average proportion of shared fragments between pairs of sampled haplotypes, as a measure of genetic diversity.

NOTE: the second approach is currently not available in version 1.0, sorry for the inconvenience!

The program also produces matrices of pairwise genetic distances between populations (with **bootstraps**) and of pairwise relatedness coefficients between individuals, and makes various tests of significance based on random permutations. It also computes the correlation between AFLP fragment sizes and frequencies when fragment sizes are provided in order to test the occurrence of size homoplasy (see Vekemans et al. 2002).

2.2. Data treated by *AFLP-SURV*

AFLP-SURV requires that the following information is provided for each individual: 1°) **name of the population** or sample to which it belongs; and 2°) its **phenotype** at each AFLP locus (**1** = fragment or marker present; **0** = fragment or marker absent; **9** = missing data, i.e. locus not scored for this individual). Dominance of the marker allele (fragment present) over the null allele (fragment absent) is always assumed.

2.3. Statistics computed

Estimates of **allelic frequencies** are computed according to either of four methods (see Krauss 2000 for a discussion):

1) Fragment frequency: the frequency of the marker (+) allele at each locus is set equal to the frequency of the AFLP fragment in the sample. This method assumes that individuals are haploids or that there is fixed homozygosity at each locus due to complete self-fertilisation. This method will give strongly biased results (large overestimation of the frequencies of the marker alleles) when applied to random mating populations with diploid genotypes.

2) The square root method (e.g. Nei 1987 p.152): the frequency of the null (-) allele at each locus is computed as the square root of the proportion of individuals in the sample that lack the AFLP fragment, i.e. the square root of one minus the frequency of the AFLP fragment in the sample. Then the frequency of the marker allele is evidently computed as one minus the frequency of the null allele. This method provides downwardly biased estimates of the frequency of the null allele, especially for loci with a high sample frequency of the AFLP fragment, i.e. low frequency of the null allele (Lynch & Milligan 1994).

3) A Bayesian method with uniform prior distribution of allele frequencies (Zhivotovsky 1999): the frequency of the null (-) allele at each locus is computed from two numbers, the sample size and the number of individuals in the sample that lack the AFLP fragment, using a Bayesian method that assumes a uniform distribution of allele frequencies. This method has been shown to efficiently reduce the bias of the square root method (Zhivotovsky 1999), however it makes the assumption that any value of the marker or null allele frequency in the range 0 to 1 is equally probable, which may not be the case in all samples.

4) A Bayesian method with non-uniform prior distribution of allele frequencies (Zhivotovsky 1999): the frequency of the null (-) allele at each locus is computed from two numbers, the sample size and the number of individuals in the sample that lack the AFLP fragment, using a Bayesian method that also estimates the distribution of allele frequencies based on the variation over loci of the frequencies of AFLP fragments in the sample. When the data concerns several populations, the distribution of allele frequencies is estimated separately for each population. This method is the more general one and is supposed to give the most accurate results (Zhivotovsky 1999), it is set as the default method. However it is computationally quite intensive, so that its use with permutation tests or with bootstraps may require overnight runs.

For methods 2) to 4) either **Hardy-Weinberg genotypic proportions** can be assumed, or a **fixed deviation from Hardy-Weinberg** genotypic proportions (specified by the user) can be enforced. In the latter case prior information on the level of inbreeding (Fis) or of self-fertilisation is required. The procedure to enforce fixed deviation from H-W is derived from Chong et al. (1994 for method 2)) and from Zhivotovsky (1999) for method 3) and 4). **No negative values of Fis are allowed** in the current version of the software, thus this procedure can only be applied when a deficit in heterozygotes is observed.

The following statistics are given at the end of the outputfile for each population and each locus: sample size (N); frequency of the AFLP fragment or marker (freq_frag); estimated frequency of the marker allele (freq_+all); estimated frequency of the null allele (freq_-all); estimated variance of the frequency of the null allele (var_-all). The relationship between these statistics and the notations of Lynch & Milligan (1994) is as follows: freq_frag = 1-x; freq_+all = p; freq_-all = q; var_-all = Var(q).

Statistics of **genetic diversity** and **population genetic structure** are computed, after estimating allele frequencies, following strictly the treatment of Lynch & Milligan (1994). Notations also follow Lynch & Milligan (1994). For precise information on these statistics, please read their paper. For each population or sample, the following statistics are computed: number (#loc_P) and proportion (PLP) of polymorphic loci at the 5% level; expected heterozygosity or Nei's gene diversity (Hj) and its variance components (total variance, Var(Hj); variance due to sampling of individuals, VarI(Hj); and variance due to sampling of loci, VarL(Hj)). The average gene diversity within populations or samples is also given (Hw) with its variance components (total variance, Var(Hw); variance due to sampling of individuals, VarI(Hw); variance due to sampling of loci, VarL(Hw); and variance due to sampling of populations, VarP(Hw)). The statistics for population genetic structure are: the total gene diversity (Ht), i.e. expected heterozygosity or gene diversity in the overall sample; the average gene diversity within populations (Hw, already discussed above); the average gene diversity among populations in excess of that observed within populations (Hb), which is analogous to Nei's Dst; and finally Wright's Fst. All these statistics are provided as unbiased estimates.

Statistics designed for pairwise **genetic distances** between populations include Nei's D (after Lynch & Milligan 1994), F_{ST}, and Reynolds' distance (-ln(1- F_{ST}) after Reynolds et al. 1983). Files with many distance matrices computed by **bootstrapping** over AFLP loci can be written optionally. These files can be directly used for instance as input for the procedures NEIGHBOR and CONSENSE from the PHYLIP software package (Felsenstein 1993) to infer bootstrap confidence on tree branches.

Estimates of the pairwise **relatedness coefficients** between individuals are computed after Lynch & Milligan (1994) and written in either matrix format for use as input in cluster or principal coordinates analyses (PHYLIP format), or in tabular format, each individual tested against each individual from each population, in a format that is convenient for use in worksheet programs to perform assignment analyses according to the distance method of Cornuet et al. (1999).

The data file is also rewritten into a file named structure.txt, in a format that can be used as input for the program STRUCTURE by Pritchard et al. (2000) in order to perform model-based clustering and automatic search for separate gene pools within a sample of individuals.

A numerical resampling procedure whereby the statistics are computed after random permutations of individuals among populations is used in order to test for significance of the Fst statistic.

If fragment sizes are provided, the Pearson correlation coefficient between fragment sizes and fragment frequencies computed on the overall sample (all populations put together) is computed, together with its significance level (P-value).

2.4. How to use *AFLP-SURV* – short overview

AFLP-SURV runs under Windows (95 or later) but has no fancy windowing features. To launch the program just double click on the program icon or on its shortcut. A single data file must contain all individual characteristics (name of population, AFLP phenotypes). Details of the analyses to be carried out (method of computation of allelic frequencies, permutation tests, bootstraps options) will be specified after the program has been launched. Most results of the analyses are written to a single output file. Data and output files are text files with **tab delimited** pieces of information. Hence they are best opened and edited using a worksheet software such as Excel.

It is possible to make the program use only a subset of the populations present in the data file for its computations, for instance if you want to compute local Fst statistics over several populations from the same region. In this case a file named here as the "subset file" should be prepared, which contains the number of populations to use (equal or smaller than the number given in the data file) and their names.

3. CREATING A DATA FILE

The data file is a text file. It is advised to create the data file using a worksheet program such as Excel and then save it as a "**tab delimited text file**".

3.1 Structure of the data file

The data file must be in the following format, with each piece of information within a line being separated by a **tab** (i.e. each piece of information put in adjacent columns if using a worksheet program to generate the data file).

- **first line**: **2 format numbers** separated by a tab in the following order, followed by the name of each locus (fragment or marker size or ID) also separated by a tab:
  - number of populations
  - number of loci (i.e. number of AFLP fragments or markers)
  - the name of each locus (e.g. Primer1_154, Primer1_425, frag1, 128, 75.4...): please avoid space characters within a name and do not use more than 20 characters per name. If you want the program to analyse the correlation between fragment sizes and fragment frequencies, please use the size of the fragment (for instance in bp.) as its name.
- **second line and next ones**: **individual data** (each line = 1 individual) grouped by populations or samples:
  - name of the population (up to 20 characters, should be **exactly the same** for all individuals from the same population or sample)
  - name of the individual (data not used by the program but can be useful for editing purpose, and for alignment of the locus names with the locus genotypes within the worksheet)
  - phenotype at each locus (either 0, 1 or 9), also separated by a tab:
    0 = fragment or marker absent in the individual
    1 = fragment or marker present in the individual
    9 = missing data, i.e. locus not scored for this individual
- **last line** (after the last individual): the word "**END**" (in uppercase)

3.2. Example of data file

This is a data file with three populations scored at four loci:

```
3       4       ACT_CAA_74      ACT_CAA_342     ACT_CTC_127     ACT_CTC_482
Tombouctou      1       1       0       0       1
Tombouctou      2       1       0       0       0
Tombouctou      3       1       1       0       0
Tombouctou      5       1       0       0       0
Tombouctou      7       1       0       0       1
Tombouctou      8       1       1       0       0
Tombouctou      10      1       0       0       1
Houyet  7c4     0       0       1       0
Houyet  7a5     0       1       1       0
Houyet  7b3     0       0       1       0
Houyet  7b5     0       9       1       1
Houyet  7a2     0       1       1       0
Houyet  7b1     0       0       9       9
Houyet  7b4     0       0       1       1
Tailfer A       1       0       0       1
Tailfer B       1       0       1       1
Tailfer F       0       0       9       1
Tailfer J3      1       0       0       1
Tailfer J4      9       0       0       1
END
```

3.3. Structure of the subset file

The subset file is an optional text file that specifies the number and the names of a subset of populations from the data file that is to be used for computations of all statistics, the remaining populations will be discarded from the analysis.

The first line of the file gives the number of populations to analyse, followed by a tab, followed by "0" (a zero).

Each of the following lines gives the name of each population to analyse. It is important that the spell of these names matches exactly the names used in the data file.

The data file must be in the following format, with each piece of information within a line being separated by a **tab** (i.e. each piece of information put in adjacent columns if using a worksheet program to generate the data file).

Here is an example of a subset file, that limits the analysis to only two of the populations from the example data file.

2 0

Houyet

Tailfer

3.4. Present data size limitations

max. 1000 populations (samples)

max. 5000 loci

max. 20 characters for the population and locus names

max. length of any line in the data file : 10 000 characters

Please contact me if these limitations are a problem for you, I may be able to send you a recompiled version with other specifications.

4. RUNNING THE PROGRAM

The program runs on PC with Windows 95 or later versions, but has no fancy windowing features. It also runs on a Macintosh under *virtual PC*.

4.1. Launching the program

Launching the program directly (without shortcut): Launch the program by double-clicking on its icon. The data file must reside in the same folder as the program file.

Launching the program through a shortcut: Using a shortcut to the program does not require that the data file resides in the same folder as the program (thus avoiding many copies of the program in different folders where different data files reside). It also allows, optionally, to specify names of the data, output and subset files before launching the program. You can create a shortcut by dragging the program file after right-clicking and choosing *CREATE SHORTCUT*. You must make sure that the shortcut points to the right folders, i.e. the one of the data file labelled as "*Start in*", and the one of the program file labelled as "*Target*". Therefore, on the shortcut sheet of the *PROPERTIES* associated to the shortcut file/object (reached by right-clicking on the shortcut file/object), check that 1°) the instruction next to *Target* gives the correct path pointing to the program file (this is normally the case when creating the shortcut), and 2°) the instruction next to *Start in* gives the correct path to the folder of the data file (usually it points to the folder of the program file so that you have to modify this instruction line). If you wish to specify the data, output and, optionally, subset file names before launching the program, write these names in the instruction line next to *Target*, just after the definition of the path pointing to the program file (e.g.: C:\programs\AFLPsurv.exe data.txt result.txt subset.txt).

4.2. Specifying the data / output / subset files

When the data / output / subset files are not specified before launching the program (see above *Launching the program through a shortcut*), you are requested to enter the names of the data, output (results) and subset files.

If you just press *RETURN* to these questions, the default names "aflpdata.txt" and "aflpout.txt" will be considered as data and output files, respectively (this can be useful if you wish to carry out many different analyses on the same data set without having to enter the file names each time), and no subset file will be used (hence all populations from the data file will be analysed).

If a file with the same name as the output file already exists in the folder, the program will ask if you wish to: erase the existing file first (enter '*e*'), add results to the end of this file (enter '*a*' or simply press *RETURN*), or change the name of the ouput file (enter the new name).

Once the program works, it first displays the basic information from the data file on the screen. The first set of information displayed is: the number of populations and their names, the number of loci, and the name of the outputfile.

If a problem has occurred when reading the data file, the program gives an **error message written to file "*error.txt*"** (this file would also contain the line numbers corresponding to the individuals for which genotype data is erroneous). An error is also displayed if the number of populations or loci found is not the one specified in the data file.

4.3. Selecting the appropriate options

You define the analyses to carry out by answering to several questions asked by the program.

"*Choose a method of computation of allelic frequencies*":

> You have to choose one of four alternative methods to estimate allelic frequencies at each locus for each population, based on the observed proportions of individuals with the AFLP fragment present (see section 2.3. for a description of the four methods). Type 1, 2, 3 or 4 accordingly, or press *RETURN* for choosing the default method 4 (Bayesian method with non-uniform prior distribution of allele frequencies). If you don't know which is best for your data, follows the chart below:

- If your organism is haploid or completely (or almost completely) self-fertilised, choose method **1**
- If it is not, then:
  - choose method **4**
  - if a run under method 4 shows up to be too slow or makes the program crash, then:
    - choose method **3**
    - if a run under method 3 shows up to be too slow or makes the program crash, then:
      - choose method **2**
      - if you are still not satisfied, buy another computer, complain to me with a copy of your data file, or surf the net for a better behaved software!

***"If you assume Hardy-Weinberg genotypic proportions, press 'Return' - If you assume deviation from Hardy-Weinberg genotypic proportions, enter Fis value"*:**

This question will arise if you choose methods 2, 3, or 4 for estimating allelic frequencies. If your organism has separate sexes or is an obligate outcrosser, you may want to assume that the genotypic proportions follow those predicted by the Hardy-Weinberg model, i.e. you assume random mating within each population.

If you have prior information pointing out to a known deviation from Hardy-Weinberg genotypic proportions, for instance based on an allozyme or microsatellite survey, enter the mean value of Wright's inbreeding coefficient (Fis) measuring this deviation. Currently, the program will only accept values of **Fis higher or equal to zero**, thus only deficit in heterozygotes is allowed (as opposed to excess in heterozygotes).

If you have prior information on the level of self-fertilisation of your organism, enter the expected equilibrium value of Wright's inbreeding coefficient (Fis) under the mixed mating model computed as Fis = s/(2-s) with s = self-fertilisation rate (e.g. Hartl & Clark 1989 p.262).

***"Enter the number of permutations for test on Fst"*:**

If you want to perform a test of significance of the overall differentiation among populations, enter the desired number of permutations. Pressing *RETURN* without typing a number will make the program skip the test. At least 100 permutations are required for testing the null hypothesis at the 5% level and 500 permutations for the 1% level. This test may take awhile when used with the Bayesian methods of estimating allelic frequencies, but you can let the program run overnight.

***"Enter the number of bootstraps for genetic distances"*:**

If you want to compute bootstrap confidence values on tree branches using a phylogenetic software such as the PHYLIP package (Felsenstein 1993), enter the desired number of bootstraps. Bootstraps are performed over loci. Three files with as many pairwise genetic distance matrices as the specified number of bootstraps will be written, one for each estimate of genetic distance (see section 2.3.). Pressing *RETURN* without typing a number will make the program skip the bootstrap computations.

Once all options have been chosen, the program will display them on the screen. It will also write them to the output file.

4.4. Information displayed during computations

Once the program proceeds to the calculations, it displays the computational stage: computation of allele frequencies, performing the permutation tests, computing the bootstraps for genetic distancss. The program can be stopped anytime by pressing "*Ctrl*" + "*c*".

When the computations are finished, the program will spontaneously close its window. You can proceed to examination of the outputfile. If the program crashed, don't forget to open the file "**error.txt**", because this may give you some information on the origin of the problem.

5. INTERPRETING THE OUTPUT FILE

All the major results are found in a single output file. The output file can be read as a text file or as an Excel worksheet; in the latter case you can change the extension *.txt* into *.xls* and open the file by double-clicking on its icon. The results appear in the following order.

Firstly, the basic information as it appeared on the screen when running the program is written: name of data file, number of populations, and number of loci.

Then the options chosen when running the program are given: method of computation of allelic frequencies; value of Fis, if relevant; number of permutations for test on Fst; number of bootstraps for genetic distances.

Follow several general statistics: total number of fragments recorded (should be the same as the number of loci); mean number of fragments present in an individual (i.e. the average number of bands per individual); total number of segregating fragments (i.e. fragments that are not always present nor always absent in all individuals) and its proportion relative to the total number of fragments.

If the locus names correspond to the sizes of the fragments (this will occur anyway if the program finds out that the locus names can be transformed to numbers) it will write the average fragment size, its standard deviation (S.D.), the Pearson correlation coefficient between fragment size and fragment frequencies computed on the overall sample (when all populations are put together), and the P-value associated with the correlation (if, say $P<0.05$, then the correlation coefficient is significantly different from zero). Size homoplasy is expected to occur when a large number of fragments is amplified for a given primer combination, and will be higher for small fragments because of the non-uniform distribution of fragment sizes under AFLP (Vekemans et al. 2002). As a consequence, a negative correlation between fragment size and frequency will be observed if size homoplasy occurs (Vekemans et al. 2002).

Follow several tables of results:

Table 1: Population data [Lynch & Milligan method]

Table 1 gives for each population the following statistics:

- name of the population
- *n* : average number of scored individuals (n)
- *#loc* : number of loci scored
- *#loc_P* : number of polymorphic loci at the 5% level, i.e. loci with allelic frequencies lying within the range 0.05 to 0.95
- *PLP* : proportion of polymorphic loci at the 5% level, expressed as a percentage
- *Hj* : expected heterozygosity under Hardy-Weinberg genotypic proportions, also called Nei's gene diversity (analogous to H or He in most publications)
- *S.E. (Hj)* : standard error of Hj
- *Var(Hj)* : variance of Hj
- *VarI(Hj)* : variance component of Hj due to sampling of individuals (finite sample size)
- *VarI%* : proportion of Var(Hj) due to sampling of individuals
- *VarL(Hj)* : variance component of Hj due to sampling of loci
- *VarL%* : proportion of Var(Hj) due to sampling of loci

Table 2: Gene diversity within populations [Lynch & Milligan method]

Table 1 gives the average, computed over all populations, of the gene diversity within populations, as well as its variance and different components of variance:

- *Hw* : the mean within-population expected heterozygosity under Hardy-Weinberg genotypic proportions, also called mean Nei's gene diversity within populations (analogous to Hs)
- *S.E. (Hw)* : standard error of Hw
- *Var(Hw)* : variance of Hw
- *VarI(Hw)* : variance component of Hw due to sampling of individuals (finite sample size)
- *VarL(Hw)* : variance component of Hw due to sampling of loci
- *VarP(Hw)* : variance component of Hw due to sampling of populations

The proportion of the total variance in Hw due to sampling of individuals, loci and populations, respectively, are also given as percentages. Please, note that the estimates of the components of variance can be negative, and thus meaningless, in several cases for unknown reasons.

The statistics for population genetic structure are: the total gene diversity (Ht), i.e. expected heterozygosity or gene diversity in the overall sample; the average gene diversity within populations (Hw, already discussed above); the average gene diversity among populations in excess of that observed within populations (Hb), which is analogous to Nei's Dst; and finally Wright's Fst. All these statistics are provided as unbiased estimates.

Table 3: Population genetic structure [Lynch & Milligan method]

Table 3 gives the statistics of population genetic structure, e.g. of genetic differentiation among populations (analogous to Nei's analysis of gene diversity).

- *Ht* : the total gene diversity
- *Hw* : the mean gene diversity within populations (analogous to Nei's Hs)
- *Hb* : the average gene diversity among populations in excess of that observed within populations (analogous to Nei's Dst), or genetic differentiation among populations
- *Fst* : Wright's fixation index, measuring the genetic correlation between pairs of genes sampled within a population relative to pairs of genes sampled within the overall set of populations (also interpreted as the proportion of the total gene diversity that occurs among as opposed to within populations).

Standard errors and variances of these statistics are also given.

Table 4: Permutation test for genetic differentiation among populations

If there are at least two populations in the overall data set and if a number of permutations higher than zero has been entered, Table 4 will provide the results of the test for genetic differentiation among populations (test on Fst). The number of permutations performed is given, as well as a reminder that individual estimates of Fst for each permutation are written to the file "**sample.txt**". I encourage you to look at the file, to better understand your data. Be careful that this file will be overwritten at each run.

The null hypothesis for the test is that there is no genetic differentiation among the populations. Values of Fst are computed after each permutation consisting in randomly permuting individuals among existing populations or samples. The set of values of Fst obtained by permutation gives an ad-hoc distribution of the statistic under the null hypothesis. The observed value of Fst (the "real" one) is then tested against this distribution.
The following statistics are produced:

- *observed*: observed value of Fst with "real" populations
- *lower 95% limit* : the value of Fst lying at the 5% leftmost part of the distribution under the null hypothesis. If the observed Fst is lower than this value, the null hypothesis is rejected as a two-sided test, and it can be concluded that the actual populations are more similar than random assemblages of the individuals (a very unlikely situation)
- *upper 95% limit* : the value of Fst lying at the 5% rightmost part of the distribution under the null hypothesis. If the observed Fst is higher than this value, the null hypothesis is rejected as a two-sided test, and it can be concluded that the actual populations are more genetically differentiated than random assemblages of the individuals.
- *lower 99% limit* : same for the 1% leftmost part of the distribution
- *upper 99% limit* : same for the 1% rightmost part of the distribution
- *P value (low)* : this gives the probability of the type I error , i.e. the probability of rejecting a true null hypothesis, as a one-sided test with the observed Fst lower than values under the null hypothesis (a very unlikely situation).
- *P value (high)* : this gives the probability of the type I error , i.e. the probability of rejecting a true null hypothesis, as a one-sided test with the observed Fst higher than values under the null hypothesis. Thus if this value is lower than, say 0.05, it can be concluded that the actual populations are more genetically differentiated than random assemblages of the individuals.

Nei_s genetic distance after Lynch and Milligan - 1994 (Phylip format)

Pairwise Fst between populations (Phylip format)

Reynolds et al. genetic distance between populations (Phylip format)

These tables provide matrices of pairwise genetic distances between each pair of populations. The format is according to the software package PHYLIP (Felsenstein 1993), such that these tables may be transferred into a text file by cut and paste and directly run with procedures such as NEIGHBOR to build a tree.

20 bootstrapped distance matrices written to files aflp_nei.txt aflp_fst.txt and aflp_reyn.txt

This sentence is to remind the user that three files with bootstrapped distance matrices have been written.

Table 5: Fragment and allelic frequencies in each population (to paste in Excel)

This final table gives for each population (set of columns) and each locus (lines) the estimates of allelic frequencies. This table is best visualised after pasting into Excel or other worksheet program.
The following statistics are given for each locus:

- *name of each locus*
- *freq_frag* : frequency of each AFLP fragment or marker in the total sample (all populations put together)

The following statistics are given for each population and each locus:

- *N* : sample size (non-missing data)
- *freq_frag* : frequency of each AFLP fragment or marker
- *freq_+all* : estimated frequency of the marker allele
- *freq_-all* : estimated frequency of the null allele
- *var_-all* : estimated variance of the frequency of the null allele

Relatedness between pairs of individuals (Phylip format)

1-r distance between pairs of individuals (Phylip format)

These tables provide matrices of pairwise genetic relatedness or distances between each pair of individuals in the overall dataset. The format is according to the software package PHYLIP (Felsenstein 1993), such that these tables may be transferred into a text file by cut and paste and directly run with procedures such as NEIGHBOR to build a tree.

Table 6: Relatedness between an individual and all other individuals grouped by population

This table writes, in turn for each individual (a) in the overall data set, the relatedness coefficient between that individual and each other individual (b) from the overall data set, grouped by population. Each line of the table reports the label of the population of individual a, the id of individual a (coded from 1 to the total number of individuals), the lable of the population of each individual b, the id of individual b, and then the relatedness coefficient between a and b (rab) and 1-rab, a measure of the genetic "divergence" between a and b. This large table can be pasted into a worksheet program such as Excel, then using the "dynamic cross-tabulation" facilities it is very easy to compute the average relatedness or distance between a given individual and each population (computed as the average relatedness between a and each individual of the population except a). According to the distance method of Cornuet et al. (1999), the population with the highest relatedness, or lower distance, is the most likely population of origin of individual a.

File structure.txt

The data file is also rewritten into a file named structure.txt, in a format that can be used as input for the program STRUCTURE by Pritchard et al. (2000) in order to perform model-based clustering and automatic search for separate gene pools within a sample of individuals. Each class of phenotype (presence or absence of the band) is treated as a separate haploid allele, with the presence of a band scored as allele 1, absence of a band scored as allele 2, and the second gene of each diploid genotype is scored as -9, the value set to missing data (see readme file from STRUCTURE, section 8.1 "Dominant loci"). Please don't forget to specify into the parameter file of STRUCTURE that the missing value code is -9 (MISSING -9) and set NOADMIX 1. Please, also set LABEL 1 and POPDATA 1.

Each individual is written onto two separate lines, the first one with the code 1,2, or -9 depending whether the fragment is present, absent, or data is missing, respectively. The second line contains only -9 because data is treated as haploid (see readme file of STRUCTURE). The first column represents the individual number (from 1 to ntot, where ntot is the number of individuals in the overall data set), that is why LABEL has to be set to 1. The second column represents the population label (coded from 1 to the total number of populations) , that is why POPDATA has to be set to 1.

6. TECHNICAL NOTES

This section will be expanded in the near future, sorry for the inconvenience!

6.1. Approach of Clark & Lanigan (1993) (currently not available)

Statistics at the population level

The program outputs for each population the average proportion of shared fragments between pairs of individuals (*Fp*), the average probability that an amplification site corresponding to the primer on the side of the high-frequency restriction enzyme which occurred in the common ancestor of a pair of individuals remains unchanged until present (*P2p*), the nucleotide diversity (*pp*).

Fp is analogous to *F* in Nei and Li (1979), but averaged over all pairs of individuals within population p, and is computed according to Clark (1997) as $\sum_i x_{pi}^2 \Big/ \sum_i x_{pi}$, where x_{pi} is the average frequency of the marker allele at locus i in population p, obtained from the observed frequency of AFLP fragment i (see section 2.3.)

P2p is analogous to P in Nei and Li (1979), but again averaged over all pairs of individuals within population p and referring to the conservation of the r2 nucleotides in the restriction site plus the s2 selective nucleotides corresponding to the high-frequency restriction enzyme. *P2p* is obtained by an iteration procedure similar to that suggested by Nei (1987)
for restriction fragment data, using the following formula (Vekemans et al., in prep.) and :

$$Fp = \frac{(P2p)^{\frac{2(r1+r2+s1+s2)}{r2+s2}}}{3-2P2p^{\frac{r2}{r2+s2}}}$$, where r indicates the number of nucleotides in the recognition sites of the low-frequency (r1) and high-frequency (r2) restriction enzymes, and s1 and s2 indicate the number of selective nucleotides in the corresponding primer sites. For regular AFLP, r1=6, r2=4, s1=s2=3 (s1 and s2 are taken from the first line of the data file).

pp is an estimate of the average nucleotide diversity π, which is the average number of nucleotide differences per site between two randomly sampled sequences (Nei, 1987), and is obtained from *P2p* as:

$pp = 1-(P2p)^{\frac{2}{r2+s2}}$. A bootstrap analysis is performed by drawing 1000 new data sets from randomly sampling AFLP fragments with replacement and resuming for each data set the calculations detailed above, and the following statistics are presented for each population: mean and standard deviation of *pp* across the 1000 bootstraps; 95% and 99% confidence intervals for the mean *pp*.

Statistics between population pairs

Then the program outputs for each pair of populations the probability of fragment sharing between populations (*F*), the average proportion of net nucleotide differences between populations (*p*), and the nucleotide divergence (*d*).

F is computed according to Clark (1997) as a ratio of the average probability that a given fragment is shared between two individuals belonging to different populations over the average probability that the same fragment is shared between two individuals belonging to the same population:

$$F = \frac{\sum_i x_{1i} x_{2i}}{\sqrt{\sum_i x_{1i}^2 \sum_i x_{2i}^2}},$$

where x_{1i} and x_{2i} refer to the average frequencies of allele i in populations 1 and 2, respectively. Note that there is a mistake in the formula given by Clark (1997, p.221) where a minus sign is given instead of a multiplication in the denominator (Clark, personal communication).

p is obtained from *F* according to the same procedure as described above for *pp*.

d is the nucleotide divergence between two populations, or number of net nucleotide substitutions between the two populations (Nei, 1987), and is obtained as:

$$d = -\tfrac{3}{4}\ln\left(1-\tfrac{4}{3}p\right)$$

A bootstrap analysis is performed by drawing 1000 new data sets from randomly sampling AFLP fragments with replacement and resuming for each data set the calculations detailed above, and the following statistics are presented for each population: mean and standard deviation of *d* across the 1000 bootstraps; 95% and 99% confidence intervals for the mean *d*.
In addition, the programs write in a file called "phy_div.txt" 1000 matrices of pairwise nucleotide divergences between populations corresponding to the 1000 bootstraps, in a phylip format that can be used to perform phylogenetic analyses with bootstrap information.

Analysis of population structure
The results of a simple analysis of genetic diversity are presented in table 4: the total nucleotide diversity (*Ht*), average nucleotide diversity within populations (*Hw*), average diversity between populations (*Hb*), and the proportion of between population diversity (*Gst*).

Ht is computed by putting all individuals in a single population and computing *pp* as described above for this overall population.

Hw is computed as a mean over populations of pp.

Hb is computed as the difference *Ht*-*Hw*; and *Gst* is computed as *Hb*/*Ht*.

7. CITED REFERENCES

Chong D.K.X., R.-C. Yang and F.C. Yeh, 1994. Nucleotide divergence between populations of trembling aspen (Populus tremuloides) estimated with RAPDs. Current Genetics, 26, 374-376.

Clark A.G., 1997. Estimating nucleotide divergence with RAPD data, pp. 219-225. In *Fingerprinting methods based on arbitrarily primed PCR*, Micheli M.R. and Bova R. (Eds.), Springer, Berlin.

Clark A.G. and C.M. Lanigan, 1993. Prospects for estimating nucleotide divergence with RAPDs. Mol. Biol. Evol., 10:1096-1111.

Cornuet J.-M., S. Piry, G. Luikart, A. Estoup and M. Solignac, 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics, 153:² 1989-2000.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington,
Seattle.

Hartl D.L. and A.G. Clark, 1989. Principles of population genetics, 2nd edition. Sinauer Associates, Sunderland, Massachusetts.

Innan H., R. Terauchi, G. Kahl, F. Tajima, 1999. A method for estimating nucleotide diversity from AFLP data. Genetics, 151, 1157-1164.

Krauss S.L., 2000. Accurate gene diversity estimates from amplified fragment length polymorphism (AFLP) markers. Molecular Ecology, 9,1241-1245.

Lynch M. and B.G. Milligan, 1994. Analysis of population genetic structure with RAPD markers. Molecular Ecology, 3:91-99.

Nei M., 1987. Molecular evolutionary genetics. Pp. 512. Columbia University Press, New York.

Nei M. and W.H. Li, 1979. Mathematical model for studying genetic variation in term of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76:5269-5273.

Pritchard J.K., Stephens M. and P. Donnelly, 2000. Inference of population structure using multilocus genotype data. Genetics, 155, 945-959.

Reynolds J., B.S. Weir and C.C. Cockerham, 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics, 105, 767-779.

Vekemans X., T. Beauwens, M. Lemaire and I. Roldan-Ruiz, 2002. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Molecular Ecology, 11, 139-151.

Zhivotovsky L.A., 1999. Estimating population structure in diploids with multilocus dominant DNA markers. Molecular Ecology, 8, 907-913.