scanAFLP v1.3 Help
2011/02/06

Authors

Doris Herrmann[1], Bénédicte N. Poncet[2] and Felix Gugerli[1]

[1] Ecological Genetics and Evolution, WSL Swiss Federal Research Institute, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland

[2] Laboratoire d'Ecologie Alpine (LECA), CNRS UMR 5553, University Joseph Fourier, BP 53, 2233 Rue de la Piscine, 38041 Grenoble Cedex 9, France

For correspondence

Bénédicte N. Poncet

e-mail: benedicte.poncet@e.ujf-grenoble.fr or felix.gugerli@wsl.ch

How to cite scanAFLP

Herrmann D., Poncet B. N., Manel S., Rioux D., Gielly L., Taberlet P., Gugerli F (2010)Selection criteria for scoring amplified fragment length polymorphisms (AFLPs) positively affect the reliability of population genetic parameter estimates. *53:302-310.*

Aim

scanAFLP implements a semi-automatic and repeatable method to score data for dominant molecular markers (e.g. AFLPs). scanAFLP is a script written and working in an R environment. This R script interprets PCR product fluorescence (fragment size and fragment height) data matrices created by softwares such as Applied Biosystems' GeneMapper and converts them into presence-absence (1-0) phenotype tables according to quality and repeatability criteria. The script is available on http://www-leca.ujf-grenoble.fr/logiciels.htm.

How to use scanAFLP

1/ Infiles

All required files are in the same folder *scanAFLP* :

- *scanAFLPv1-0.r*
- *infile.txt*
- *parameters.txt*
- *controls.txt*

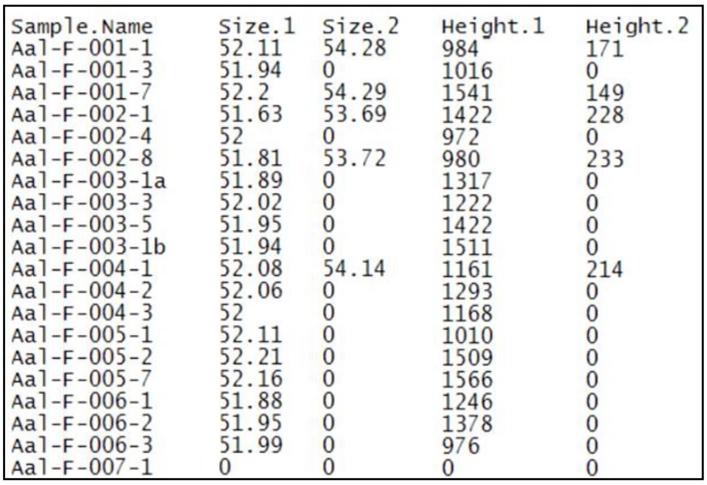
scanAFLPv1-0.r is the R script where the scoring procedure is implemented. The steps of the selection of markers are briefly introduced here, but all details are available in Herrmann *et al.* (2010):

- step A importation of the data infile from a preliminary manual non-restrictive selection of markers on reading electropherogram softwares. PCR fragments with a height lower than the ThresholdA1 (in relative fluorescence units, rfu) are set to 0, i.e. fragments are assessed as absent. Each marker will be labeled using the primer combination name followed by the average size of fragments of the respective marker. Markers without any remaining fragments are removed. Controls are removed except the first replicated sample in all replications (duplicated samples and multiple controls). The matrix is transformed into a binary matrix to obtain the initial matrix A. It is only the binary matrix of presence-absence of markers resulting from a software such as GeneMapper.

- step B restrictive selection of high-quality markers and removal of noise. PCR fragments with a height lower than the absolute threshold TresholdB1 (more stringent than TresholdA1) are set to 0. Markers that have a majority of small peak heights are removed. Next, fragments with a height lower than a given percentage PctgB2 of the mean height of the maximum height frequency class are set to 0. Finally, the coefficient of variation (CV = standard deviation / mean) of fragment height of each marker is calculated and markers with a CV higher than CVB3 are excluded. Markers without any remaining fragments are removed. Controls are removed except the first replicated sample in all replications (duplicated samples and multiple controls). The matrix is transformed into a binary matrix to obtain the initial matrix B.
After each step, markers with only one fragment with a fragment height lower than the threshold TresholdB4 are removed.

- step C exclusion of markers with low repeatability. Repeatability for controls among plates and within plates is calculated.
A marker is excluded if more than NC1 differences occur between all pairs of samples analyzed twice (controls within plates) or, if more than NC2 differences occur among the repetitions of controls among plates or, if this marker is present in almost NC3 negative controls. Markers without remaining fragments are removed. Controls are removed except the first replicated sample in all replications (duplicated samples and multiple controls). The matrix is transformed into a binary matrix to obtain the initial matrix C.

infile.txt is the data matrix resulting from electropherogram evaluation obtained from softwares such as Applied Biosystems' GeneMapper. Alternative softwares could also be used to easily obtain similar infiles. You have to specify its name so this file does not need to be called *infile.txt*. Input files should be in a text (tab delimited) format (.txt) and contain a first column with sample names, all columns of peak sizes and finally all columns of non-normalized peak heights in the same respective order (Fig.1). Peak size is the estimated length of a single PCR fragment or band within an AFLP profile, and peak height is the maximum fluorescence intensity of a single PCR fragment or band within an AFLP profile. An infile example, *infile_example.txt*, is provided. We recommend not to include any spaces in the row or column names for the input table, and missing values are not allowed.



```
Sample.Name     Size.1  Size.2   Height.1   Height.2
Aal-F-001-1     52.11   54.28    984        171
Aal-F-001-3     51.94   0        1016       0
Aal-F-001-7     52.2    54.29    1541       149
Aal-F-002-1     51.63   53.69    1422       228
Aal-F-002-4     52      0        972        0
Aal-F-002-8     51.81   53.72    980        233
Aal-F-003-1a    51.89   0        1317       0
Aal-F-003-3     52.02   0        1222       0
Aal-F-003-5     51.95   0        1422       0
Aal-F-003-1b    51.94   0        1511       0
Aal-F-004-1     52.08   54.14    1161       214
Aal-F-004-2     52.06   0        1293       0
Aal-F-004-3     52      0        1168       0
Aal-F-005-1     52.11   0        1010       0
Aal-F-005-2     52.21   0        1509       0
Aal-F-005-7     52.16   0        1566       0
Aal-F-006-1     51.88   0        1246       0
Aal-F-006-2     51.95   0        1378       0
Aal-F-006-3     51.99   0        976        0
Aal-F-007-1     0       0        0          0
```

Figure 1. Example of infile with 20 samples (rows) and 2 AFLP markers (2 columns of peak sizes and 2 columns of peak heights).

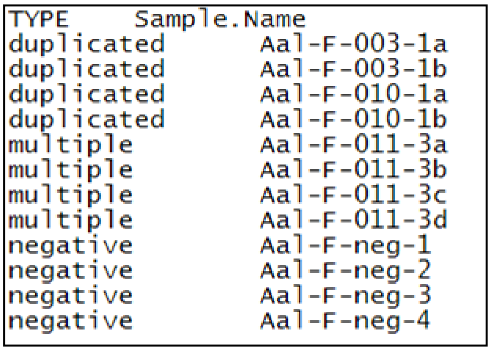
parameters.txt is a text file setting quality and repeatability thresholds to select dominant markers. Given selection parameters are based on the authors' experience with AFLP scoring, while they can be individually adjusted in the script according to particular requirements. The file contains a column of parameter names and a column of the parameter values (Table 1).

Table 1. Parameters required in scanAFLP and to specified in the file parameters.txt

PARAMETER	EXAMPLE VALUE	EXPLANATION
Species	Aal	name of the model species or an experiment code
Primers	EAATMCAC	primer combination code
Duplicated_samples	2	number of duplicated samples (controls within plates)
Multiple_controls	1	number of replicated samples (controls between plates, multiple controls)
Negative_controls	4	number of negative controls
TresholdA1	50	threshold for minimal peak height in GeneMapper (in rfu)
TresholdB1	300	absolute threshold for minimal peak height (in rfu)
PctgB2	0.1	relative threshold for minimal peak height (percentage of the mean peak height)
CVB3	1	maximal coefficient of variation of peak heights
TresholdB4	300	minimal peak height for marker with a unique peak (in rfu)
NC1	1	number of errors allowed among all duplicated samples to keep a marker

NC2	2	number of errors allowed among all replicated samples to keep a marker
NC3	1	number of presences allowed among all negative controls to keep a marker

controls.txt is a text file with a column of control types (duplicated, multiple or negative controls) and a column of corresponding sample names (Fig. 2). Controls may have the same name. If a type of control is absent, set its value to 0.



```
TYPE     Sample.Name
duplicated       Aal-F-003-1a
duplicated       Aal-F-003-1b
duplicated       Aal-F-010-1a
duplicated       Aal-F-010-1b
multiple         Aal-F-011-3a
multiple         Aal-F-011-3b
multiple         Aal-F-011-3c
multiple         Aal-F-011-3d
negative         Aal-F-neg-1
negative         Aal-F-neg-2
negative         Aal-F-neg-3
negative         Aal-F-neg-4
```

Figure 2. Example of controls.txt file with 2 duplicated samples, 1 multiple control, and, 4 negative controls.

2/ R installation and use of scanAFLP in R

a. Install the R statistical computing software from www.r-project.org.
b. Download the folder scanAFLP available at www-leca.ujf-grenoble.fr/logiciels.htm.
c. Open R.
d. Change the working directory of R to the scanAFLP directory specifying its pathway.
e. Complete parameter values in *parameters.txt*, and check that the files *infile.txt* and *controls.txt* are correct.
f. Source the scanAFLP script in R writing the following command: source('scanAFLPv1-0.r') and press enter.
g. Apply scanAFLP on your data set writing the following command: scanAFLP('infile.txt') and press enter.
h. Wait that R performs all analyses.

3/ Outfiles

All output files produced are saved in the working directory.
Three matrices A, B and C are produced. They are text files (tab delimited) giving the phenotype of all samples (rows) for each marker retained, i.e. the presence (1) or absence (0) of a single PCR fragment or band (columns). Only one of the replicated samples (controls) is randomly kept. Marker names follow the structure Species_Primers_MeanPeakSize.
In addition, the R script creates a summary text file reporting the global mismatch error, names of excluded markers and information about remaining markers after each step of the selection procedure described (Table 2).

Table 2. Description on markers available in the log.txt file at each step of the selection procedure

Code	EXPLANATION
NbSamples	number of samples
NbSamplesWithNA	number of samples with missing values
NbMarkers	number of markers
PrevMin	minimal prevalence of markers (frequency of presence)
PrevMax	maximal prevalence of markers (frequency of presence)
PrevMean	mean prevalence of markers (frequency of presence)
NbMono	number of monomorphic markers
NbPoly	number of polymorphic markers
LengthMin	minimal size of marker (base pair)
LengthMax	maximal size of marker (base pair)
LengthMean	mean size of marker (base pair)
PrevLengthCorr	Pearson's product moment correlation coefficient between sizes of markers and their prevalence
PrevLengthCorrPvalue	corresponding *P* value
MinNbPeak	Minimum number of peaks in a sample
MaxNbPeak	Maximum number of peaks in a sample
ProfilesWithoutPeaks	Number of samples without present marker (failure of PCR)
MeanNbPeak	Mean number of peaks in samples

Bugs, suggestions

Please report any bugs or errant behavior of scanAFLPv1.3 to Bénédicte N. Poncet. If you have any questions, comments or suggestions, please also contact Bénédicte.