Automatize scoring of AFLP datasets with RawGeno: a free R CRAN library.

Nils ARRIGO, Dorothée EHRICH and Nadir ALVAREZ.

1 Getting started

1.1 Overview of the analysis

Analysing AFLP electropherograms is achieved using two programs: PeakScanner and RawGeno. Whereas PeakScanner detects AFLP peaks along electropherograms and calculates their intensity and size (by relying on an internal size standard included in electropherosis), RawGeno proceeds to the binning and scoring of AFLP electropherograms.

RawGeno includes several filters to assess the quality of electropherograms and checks the consistency of binning. In addition, several preliminary analyses are available to the user for making biological inferences and/or remove outlier samples. Finally, several functions allow exporting resulting data sets into properly formatted files for further analyses.

1.2 Organizing your project

In RawGeno 2.0 the AFLP scoring project should be organized according to the following procedure.

1. Create a folder (hereafter “project folder”) from which the project will be managed.
2. In this folder, add a sub-directory including all electropherogram files (*.fsa). Also add an R shortcut (Windows users) to conveniently launch RawGeno scoring sessions. Right-clicking on this R shortcut allows defining the default working directory of R by specifying it into the “Start into” addressing field of the shortcut. Copy-pasting the project folder address into this field will set-up the working directory of R accordingly.

Create a text-tabulated table censing individuals included in the project (hereafter referred to as “info table”). The info table is optional as the minimal RawGeno analysis can proceed without it. However, RawGeno includes several functions relying on this table, for instance to label individuals during preliminary analyses or facilitate the sorting and selection of individuals (for example, according to populations or species) during the production of exports.

Therefore, the info table should include any additional relevant information that the user would like to consider. It must contain at least the name of individuals (i.e. in a column named “Tag”) and any supplementary information in extra columns.

Tag	**Plate**	**Pos**	Species	DNA_quality
BEv14-2	1	A1	BEv	0
BEv2-2	1	A2	BEv	0
BF29-2	1	A3	BF2	0
BF36-2	1	A4	BF3	Low
BF36-2b	1	A5	BF3	Low
BHa1-2	1	A6	BHa	0
BHa3-2	1	A7	BHa	0
…	…	…	…	…

Fig. 1 Example of info table. The Tag column is mandatory (in red), Plate and Pos (in green, PCR plate and well name) are optional. In black: any relevant data to be compared with AFLP data. Pay attention to avoid weird characters (such as #, @ and others).

1.3 Installing the softwares

RawGeno works in combination with PeakScanner, an electropherogram analyser freely distributed by ABI.

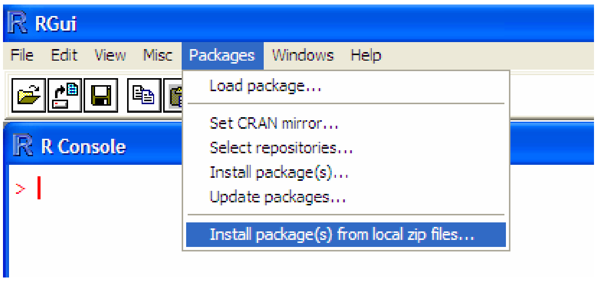


Fig. 2 Installing RawGeno. Using the graphical interface of R CRAN. RawGeno is provided as a zip file that has to be unpacked into the "library" folder of R. This operation can also be done manually with your files explorer. Similarly, uninstalling RawGeno is done by removing it from the "library" folder of R.

1.3.1 Installing PeakScanner

PeakScanner is freely available from http://marketing.appliedbiosystems.com/mk/get/PS1_login. This is a windows distribution. Linux users might run it through Wine (for the command version at least, refer to PeakScanner documentation).

1.3.2 Installing RawGeno

RawGeno is a library of R CRAN, which is freely available from http://cran.r-project.org. Of course, you must first install R CRAN before willing to use RawGeno.

RawGeno is freely available from http://sourceforge.net/projects/rawgeno as a zip file.

In windows, the installation is achieved either using the graphical user interface of R (menu "Packages/Install package(s) from local zip files") or the following command line in the R console:

```
utils:::menuInstallLocal()
```

Installing RawGeno with Linux requires to decompress RawGeno.zip into the library folder of R (see Note 1). Using the shell command line, this is done as follows (linux command lines, starting from the folder where RawGeno.zip was downloaded):

```
sudo mv RawGeno.zip /usr/lib64/R/library/RawGeno.zip
cd /usr/lib64/R/library/RawGeno.zip
sudo unzip RawGeno.zip
sudo rm RawGeno.zip
```

Finally, RawGeno requires the installation of two companion packages: vegan and rpanel, that both are available from usual R CRAN repositories. Their installation is achieved either using the graphical user interface of R (menu "Packages/Install package(s) from CRAN") or with the following command lines (prompted into the R console):

```
install.packages("vegan")
install.packages("rpanel")
```

Note 1. Linux users might need to run R as "sudo" users to properly install companion packages (i.e. vegan and rpanel). In addition, troubles might arise because R libraries are downloaded as source code and compiled locally before being installed. This requires that all compilers needed by R (such as gc, gcc, gcc-fortran and others) have been installed locally, before attempting the installation of external R packages. In OpenSUSE, the necessary compilers can be obtained using YaST2 (into the rpm groups dedicated to development tools). Ubuntu users are more fortunate because Synaptic Manager can install ready-to-use R libraries in addition to usual compilers (refer to http://cran.r-project.org/bin/linux/ubuntu/README for further details regarding repository addresses).

1.4 First steps with RawGeno

RawGeno can be launched by copy-pasting the following commands in the command line window of R:

```
require(RawGeno)
require(vegan)
RawGeno()
```

RawGeno is interfaced in a way to “guide” users from the importation of data to the export of final results. The graphical user interface (hereafter “GUI”) appears with four main menus : 1. Files, 2. Scoring, 3. Quality check, 4. Save.

1.4.1 The Files menu

Gives access to all importation steps required for the analysis. More specifically:

- Files / Electroph. / PeakScanner (*.txt): launches the importation device of PeakScanner results. To be used to analyse the electropherograms of interest.
- Files / Import / Single datasets (Binary Table): imports datasets that were already scored (and stored as simple binary tables, with lines and columns being individuals and AFLP loci, respectively). This importation function is helpful for users willing to visualise their data within RawGeno or produce exports towards external statistical programs.
- Files / Import / Merge several datasets (Binary Tables): imports batches of datasets that were already scored and merges them into a single binary matrix. This is useful for merging results from several AFLP primer pairs.

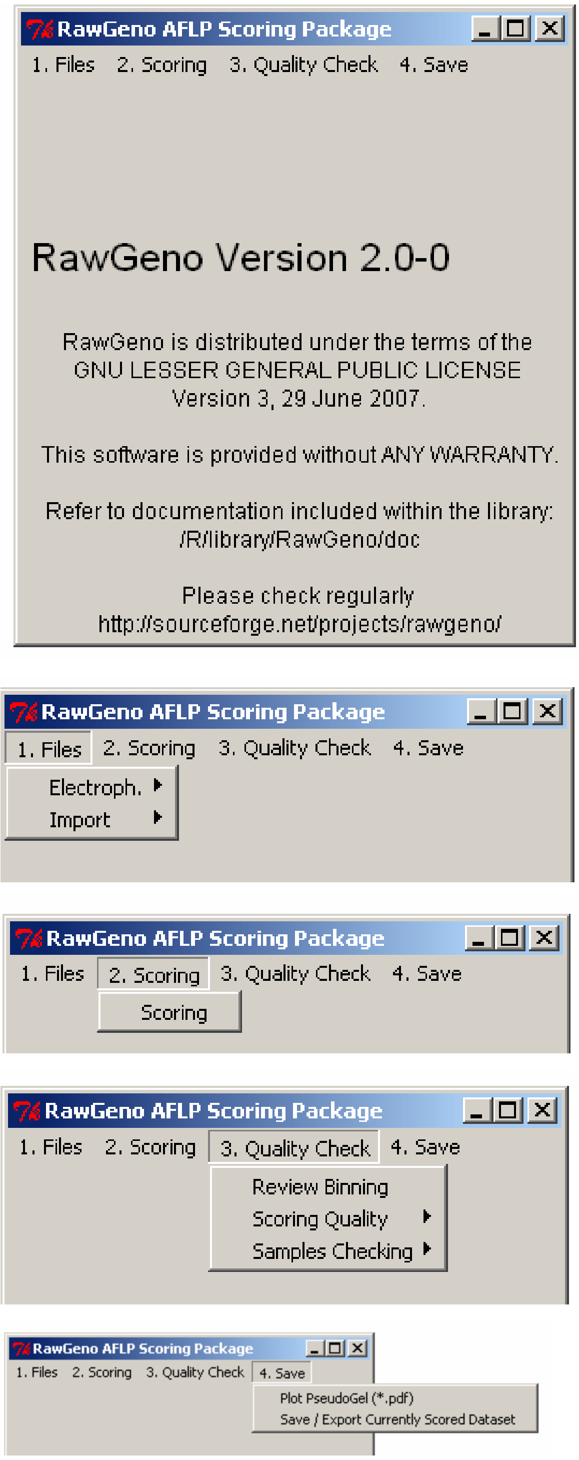
1.4.2 The Scoring menu

To be used once data from electropherograms have been imported into RawGeno.

- Scoring / Scoring: for… scoring data from electropherograms, once they have been successfully imported within RawGeno. This launches a GUI where parameters of the binning algorithm and the filtering of bins can be managed.

1.4.3 The Quality check menu

To be used once electropherograms have been scored.



RawGeno Version 2.0-0

RawGeno is distributed under the terms of the
GNU LESSER GENERAL PUBLIC LICENSE
Version 3, 29 June 2007.

This software is provided without ANY WARRANTY.

Refer to documentation included within the library:
/R/library/RawGeno/doc

Please check regularly
http://sourceforge.net/projects/rawgeno/

Fig. 3 RawGeno main interface. In addition, consider the R CRAN console where messages / statistics are displayed during the analysis.

- Quality Check / Review Binning: because binning and filtering are done by algorithms, one might want to have a look at what happened to the data. This menu launches a GUI where users can see how bins were defined along the electropherograms. This device also allows manual editing of the binning.
- Quality Check / Scoring Quality: provides statistics regarding the dataset and bins properties. More specifically:
    - Binning diagnostics: shows various statistics about the quality of bins, along the electropherograms or in general (i.e. as scatterplots or distributions).

  - Scoring statistics: provides standard statistics regarding the dataset (i.e. number of bins that were filtered during the scoring, etc.)
  - Scoring parameters: a summary of parameters that were used during the scoring.
- Quality Check / Samples Checking: my preferred! It allows user to explore their results. More specifically:
  - Visualise samples: launches a GUI where heatmap and principal coordinate analyses can be performed on the AFLP dataset.
  - Export Sample Diagnostic Values: produces a table where quality statistics regarding samples are saved.
  - PCR Plates Check: launches a GUI for visualizing how AFLP reactions performed across PCR plates. This is especially useful during wet-lab optimizations.

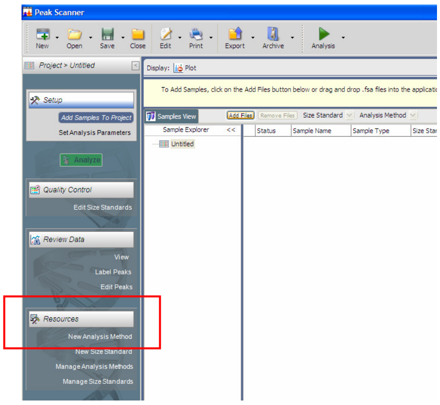
1.4.4 The Save menu

To be used once electropherograms have been scored and quality checked

OR when willing to save results from datasets that were already scored.

- Plot PseudoGel (*.pdf): saves the complete scoring project as a “gel-like” image. Still experimental
- Save / Export Currently Scored Dataset: launches a GUI for managing the export of results towards various formats.

2 Analysing electropherograms with PeakScanner



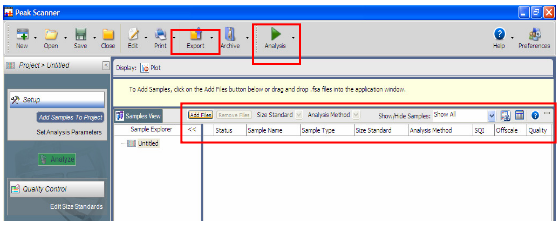
To Add Samples, click on the Add Files button below or drag and drop .fsa files into the applicati

The analysis of AFLPs starts by using PeakScanner in order to detect peaks along electropherograms and to calculate their size. The procedure is highly automated, leaving the user to set peak detection parameters and check the quality of electropherograms.

The peak detection parameters are set up using a so-called “Analysis Method”, which is available from the graphical interface (menu “Resources/Manage Analysis Methods”). Typically, a proper peak detection attempts to detect only peaks that are biologically relevant and exclude peaks only reflecting technical background noise.

We advise to set the “Analysis Method” using the following guidelines.

- Prior to the detection of peaks *per se*, a light smoothing of electropherograms might be desirable, in order to eliminate small secondary peaks inherent to technical background noise.
- The detection of peaks is achieved through a “sliding window” analysis that inspects electropherograms locally. Within the inspected region, PeakScanner first creates a modelled version of the electropherogram by fitting a polynomial curve to the data. Peaks are detected according to this modelled signal, based on their absolute width. Therefore, the detection sensitivity is adjusted by modifying the width of the sliding window (i.e. in terms of data points, the smaller it is, the more sensitive the procedure becomes), the goodness of fit reachable by the polynomial curve (again, increasing the polynomial degree of fitting increases the detection sensitivity) and the minimal width above which a peak is recorded as present. We advise to use default parameters as a starting point, as they have been shown to provide reliable results ***(4, 5)***: set 15 points for the sliding window width, use a third degree polynomial curve and consider peaks that at least have two points of half-width.
- Downstream to peak detection, PeakScanner filters peaks according to their absolute fluorescence intensity, i.e. the peak height, measured in relative fluorescent units (rfu). Visually checking electropherograms obtained from blank samples generally helps to adjust the fluorescence threshold to the upper limit of the technical background noise. While some applications might benefit from considering only peaks with a strong fluorescence (e.g. greater than 150 rfu to provide conservative estimates for band presence statistics), most users will prefer using a more permissive threshold at this stage and apply *a posteriori* filtering strategies based on bin quality statistics ***(1 - 4)***. We advise to use 50 rfu as a minimal fluorescence for considering individual AFLP peaks.
- Save the customized “Analysis Method” in order to use it during electropherogram analysis.



To Add Samples, click on the Add Files button below or drag and drop .fsa files into the application window.

Once the “Analysis Method” is set up, import the electropherograms (stored as *.fsa files) into PeakScanner using the “Add Files” button.

Define the size standard and the “Analysis Method” to be used for all individuals included into the project (set this information for the first individual, then select the columns “Size Standards” and “Analysis Method” and use the “ctrl+D” keyboard shortcut to apply these settings to the remaining individuals).

The detection and sizing of peaks is processed using the “Analysis” menu, from the graphical interface. Once achieved, electropherograms can be visualized and compared among individuals. This might help identifying AFLP reactions that were not successful (e.g. individuals with a systematically low fluorescence or showing abnormal peaks). Removing such individuals prior to the RawGeno analysis will help to enhance the final quality of scoring.

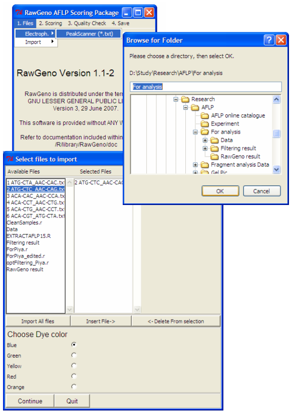
The PeakScanner analysis ends with a simple export process in which the list of peaks detected throughout the complete set of analyzed individuals is stored in a table. This is achieved using the “Export/Export Combined Table” menu, producing a text-tabulated file containing the size, height, area and width of all detected peaks (this can be checked using the “Edit Table Settings” menu).

3 Importing data within RawGeno

3.1 Importation

Importing the PeakScanner text-tabulated file in RawGeno is done using the menu

Files/RawGeno/Electroph./PeakScanner (*.txt).


RawGeno AFLP Scoring Package

RawGeno Version 1.1-2

RawGeno is distributed under the ter GNU LESSER GENERAL PUBLIC LI Version 3, 29 June 2007.

This software is provided without ANY W

Refer to documentation included within /R/library/RawGeno/doc

Browse for Folder

Please choose a directory, then select OK.

D:\Study\Research\AFLP\For analysis

For analysis

Select files to import

Available Files: 1 ATG-CTA_AAC-CAC.tx, 2 ATG-CTC_AAC-CAG.tx, 3 ACA-CAC_AAC-CCA.tx, 4 ACA-CCT_AAC-CTG.tx, 5 ACA-CTG_AAC-CCT.tx, 6 ACA-CGT_ATG-CTA.txt, CleanSamples.r, Data, EXTRACTAFLP15.R, Filtering result, ForPiya.r, ForPiya_edited.r, optFiltering_Piya.r, RawGeno result

Selected Files: 2 ATG-CTC_AAC-CAC

Choose Dye color

Blue ●
Green ○
Yellow ○
Red ○
Orange ○

During importation, RawGeno handles a single dye color at a time, which is user-specified and considers the “dye” parameter with the following values: "B" (blue; FAM), "G" (green; HEX), "Y" (yellow; NED), “R” (red; ROX) or “O” (orange; LIZ). If electropherosis was achieved using several dyes simultaneously (e.g. multiplexing of PCR products), each dye must be analysed separately in RawGeno. Datasets obtained from several dyes can be merged *a posteriori* in a final binary table (see below).

Note 2. Importing data from other sequencers. RawGeno device handles text-tabulated files produced by PeakScanner. However, results from other genescan systems can be inputted within RawGeno by following an ad-hoc data preparation. More specifically, RawGeno imports a table where all peaks detected in the dataset are stored with the following information:

Dye/Sample Peak	Sample File Name	Size	Height	Area in BP
B, 1	BEv14-2.fsa	50.9319	5943	4972
B, 2	BEv14-2.fsa	53.7099	9291	7676
B, 3	BEv14-2.fsa	56.0962	69	18
B, 4	BEv14-2.fsa	57.0999	249	91
B, 5	BEv14-2.fsa	61.7942	7334	4659
B, 6	BEv14-2.fsa	63.2218	353	259
B, 7	BEv14-2.fsa	64.6425	632	418
B, 8	BEv14-2.fsa	65.792	384	286
…	…	…	…	…

- Dye/Sample Peak = color code (B, G, Y, O or R), Peak number
- Sample File Name = sample where the peak was observed (the *.fsa is optional)
- Size = peak size (in bp)
- Height = peak fluorescence (as measured by the sequencer)
- Area in BP = peak area (optional, required by the binning edition device of RawGeno)

3.2 Filtering of low quality samples

Because the detection of AFLP peaks is based on a defined threshold, it is not easy to handle reactions showing electropherograms with varying intensities.

When improperly handled, such a situation leads to the inclusion of samples characterised by many false-absences in the final dataset.

Although the only way to correctly address this issue is a robust wet-lab protocol, RawGeno still attempts limiting the influence of low quality AFLPs on binning and scoring, by filtering individuals that were unsuccessful before proceeding to binning.

Here, the variability in the number of peaks detected per individual is used as a proxy of AFLP reactions quality. Empirical evidence shows that this statistic is dependent of the specific dataset used and the biological organism studied (see Note 3). The lower bound of this distribution most generally includes individuals with low AFLP intensities, characterized by many AFLP peaks that remain undetected in the electropherograms. Because such individuals usually represent a small fraction of the complete project, we advise removing them from the dataset. The upper bound of the distribution can either reflect a biologically relevant signal (e.g. hybridization) or a technical bias (e.g. contamination, odd PCR reaction). Such individuals should be either discarded or identified as outliers for proper interpretation in further analyses.

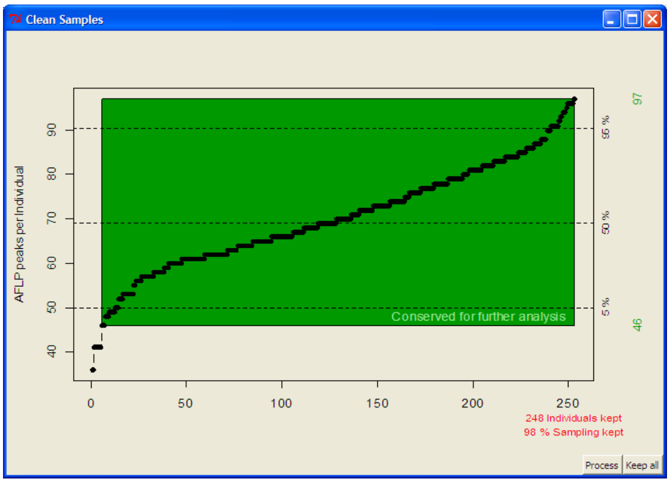


Fig 4. Filtering low quality samples. Samples are sorted according to their number of AFLP peaks. Click-and-dragging the green area allows selecting samples to be conserved for further analyses.

Note 3. Datasets with multiple species. Because the number of peaks per sample is species-dependent, users should expect a multimodal curve if several species are present in the dataset.

4 Scoring

Scoring is launched using the menu

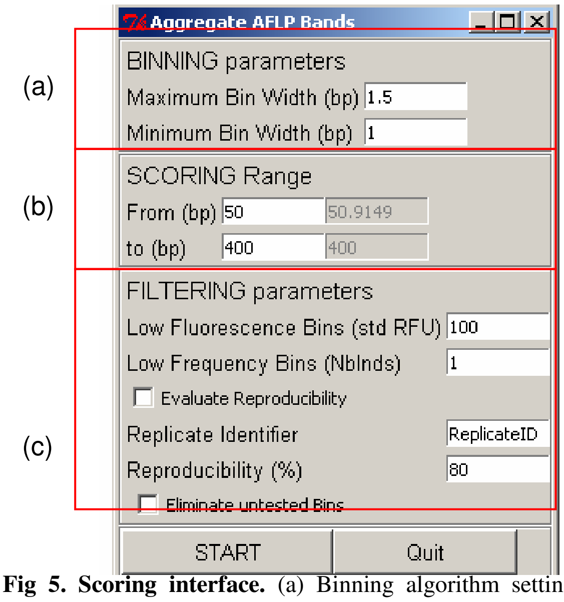
Scoring / Scoring

4.1 Binning

Building a presence/absence matrix requires recognizing which AFLP peaks are homologous across individuals. This procedure relies on the size of peaks along electropherograms and assumes homology for peaks sharing identical sizes. Because peak sizes are determined empirically using electropherosis, measurements generally include technical variations preventing the observation of identical sizes across homologous AFLP peaks. Indeed, Holland *et al.* ***(5)*** reported measurement variations ranging between 0 and 0.66 bp (with 0.08 bp in average) for replicated AFLP peaks. Therefore, properly recording the signals of AFLP peaks asks for taking into account size variations by defining size categories (i.e. “bins”) into which the presence / absence of AFLP peaks are recorded. Bins are characterised by their position along the electropherogram (i.e. the average size of peaks they include) and their width (the size difference between the largest and shortest peak included in the bin).

RawGeno uses a binning algorithm relying on the size of AFLP peaks over all individuals included in the project and defines bins in a way that respects the two following conditions.

- The first constrain is a maximal bin width. This prevents the definition of too large bins that could lead to homoplasy (i.e. erroneously assigning non-homologous AFLP peaks within the same bin). This limit is set using the MaxBin parameter. **We advise to use MaxBin values ranging between 1.5 and 2 bp** (from our experiments). Using small values (i.e. MaxBin < 0.5 bp) should be avoided as this generally causes oversplitting, a situation where the presence / absence of homologous AFLP bands are coded using an exaggerated number of bins.
- The second constrain prevents the assignment of more than one peak from the same individual within the same bin [i.e. “technical homoplasy” ***(1)***]. If such a situation occurs, RawGeno defines two separate bins in which the two peaks are assigned. This constrain can be relaxed increasing the “MinBin” parameter in order to include two peaks of the same individual in the same bin (when not exceeding the “MinBin” value in size difference). Such a relaxing might be desirable, for instance when artefactual peaks (i.e. shoulder, stutter or secondary peaks bordering the authentic peak in a same individual) lead to the definition of numerous extra-bins. In such a situation, artefactual peaks can cause the local definition of extra bins into which homologous peaks can be inconsistently assigned. **We advise to use MinBin values ranging between 1 and 1.5 bp.**



BINNING parameters

Maximum Bin Width (bp) 1.5

Minimum Bin Width (bp) 1

SCORING Range

From (bp) 50 50.9149

to (bp) 400 400

FILTERING parameters

Low Fluorescence Bins (std RFU) 100

Low Frequency Bins (NbInds) 1

□ Evaluate Reproducibility

Replicate Identifier ReplicateID

Reproducibility (%) 80

□ Eliminate untested Bins

Fig 5. Scoring interface. (a) Binning algorithm settings (i.e. define how bins are defined), (b) Scoring range (define the electropherogram region over which the scoring is performed), (c) Post-scoring filtering (removal of singletons and bins with poor fluorescence or reproducibility).

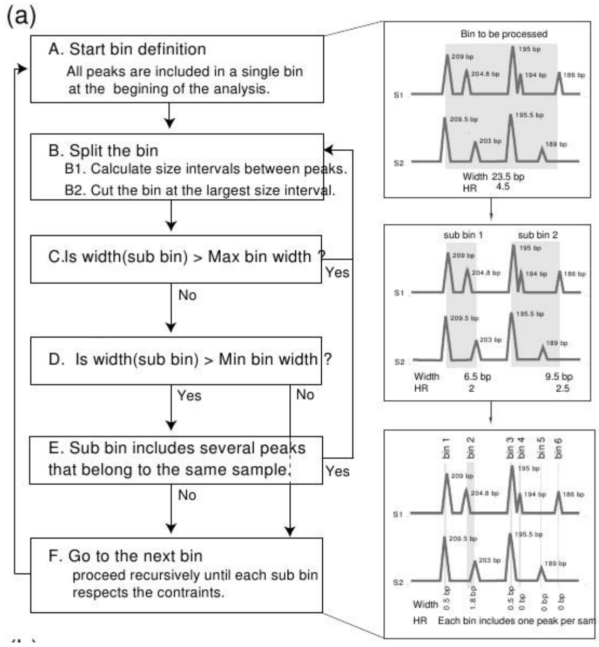

(a)
A. Start bin definition
All peaks are included in a single bin at the begining of the analysis.
B. Split the bin
B1. Calculate size intervals between peaks.
B2. Cut the bin at the largest size interval.
C.Is width(sub bin) > Max bin width ?
Yes
No
D. Is width(sub bin) > Min bin width ?
Yes
No
E. Sub bin includes several peaks that belong to the same sample?
Yes
No
F. Go to the next bin
proceed recursively until each sub bin respects the contraints.
Bin to be processed
S1
S2
209 bp
204.8 bp
195 bp
194 bp
186 bp
209.5 bp
203 bp
195.5 bp
189 bp
Width 23.5 bp
HR 4.5
sub bin 1
sub bin 2
Width 6.5 bp 9.5 bp
HR 2 2.5
bin 1
bin 2
bin 3
bin 4
bin 5
bin 6
Width 0.5 bp 1.8 bp 0.5 bp 0 bp 0 bp 0 bp
HR Each bin includes one peak per sam

Fig 6. Binning algorithm implemented in RawGeno. Left panel: main steps followed by the algorithm to define bins; right panel: illutration of binning with two samples (S1 and S2); the bin widths (i.e. the difference in size between the largest and the shortest amplicons included in the considered bin) and the technical homoplasy rates (i.e. HR, the mean number of peaks belonging to the same sample that are included in a same bin and Width, the size difference between the shortest and largest peak included into the bin).

4.2 Filtering

Once defined, bins can be filtered according to their properties and/or quality. Note that such filtering strategies require analyzing AFLP reactions with a consistent quality across individuals. In its current version, RawGeno includes three kinds of filters.

- The size filter restricts binning to a given portion of the electropherogram (i.e. the "scoring range"). We advise to limit the binning to peaks included in the range of the size ladder, because their size is accurately interpolated by PeakScanner (in contrast to larger peaks where the size is extrapolated). We recommend discarding peaks with small sizes (i.e. smaller than 100 bp, RMIN=100) as they are more likely to be homoplasic ***(6, 7)***.
- The second filter eliminates bins according to their average fluorescence. This filter assumes that bins with a high average fluorescence retrieve a more consistent signal than bins with a low fluorescence. The rationale for this strategy is the following. The fluorescence of an AFLP fragment largely determines its detection probability during the PeakScanner analysis of electropherograms. Therefore, fragments that systematically produce low fluorescences are more likely to be erroneously censed as absent from electropherograms as they might pass the threshold in some reactions but not in others just by chance.
- The reproducibility filter evaluates bin quality according to their robustness across AFLP reactions, by relying on replicated samples. This filter assumes that replicated individuals were selected randomly from the original dataset, in a way to scan the genetic diversity at best. Consider that RawGeno identifies replicated individuals using their names. Replicates must be named using the original individual name plus a suffix letter. The suffix is matched using the "who" parameter of the filtering algorithm. As an example, "mysample.fsa" and "mysampleB.fsa" are a pair of original-replicated samples, being identified with a "B" suffix (therefore, set who = "B" when filtering). For each bin, RawGeno compares original to replicated individuals and calculates the percentage of original-replicated pairs for which the AFLP signal is successfully reproduced. Bins where reproducibility cannot reach a satisfactory rate are eliminated from the final dataset.

Note 4. managing replicated samples. The replicated sample must have the same file name as the original sample, with a distinctive suffix character (the "ReplicateID").

Original samples: "E27-1.fsa", "F26-1.fsa", "F28-1.fsa"

Replicates : "E27-1x.fsa", "F26-1x.fsa, "F28-1x.fsa"

ReplicateID : x

You might want using file-managing programs such as "Ant renamer" http://www.antp.be/software/renamer.

IMPORTANT: not all bins can be checked for reproducibility, if the dataset is not completely replicated. In such a situation, users can choose to either eliminate the untestable bins, or to keep them included in their final dataset.

5 Quality checking

5.1 Binning quality

5.1.1 Manual reviewing

Binning is an automated and straightforward analysis step that users might want to review interactively. RawGeno includes a visualization device for manually editing the binning by adding, removing or modifying the width and position of bins. This device includes several help-to-decision statistics such as the average size and the number of presences associated to each bin.

5.1.2 Quality statistics

RawGeno provides statistics about the dataset quality, from the viewpoint of bins.

- Bin width, the difference between the largest and shortest AFLP bands included into the focal bin.

- Technical homoplasy, the average number of peaks belonging to same individual, included into the same bin.

- Size homoplasy, according to Vekemans et al. ***(6)***, size homoplasy can be detected by measuring and testing the linear correlation existing between the size of bins and their frequency. This procedure is already implemented in the program AFLPsurv. A negative and significant correlation suggests the occurrence of size homoplasy in the dataset.

- Peaks intensity, the fluorescence of AFLP peaks.

These values are displayed either according to the bin size (i.e. the position along the electropherogram) or as distributions.

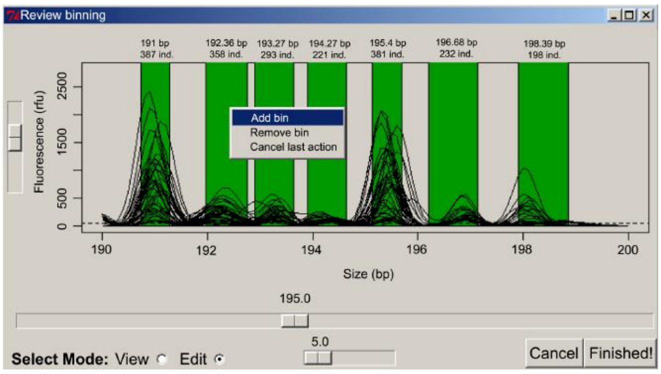


Fig 7. Binning review interface. Allowing users to shift, resize, add or remove bins interactively, starting from bins that were initially defined by the automated algorithm. Summary statistics (i.e. the average size of bins and the number of present AFLP peaks per bin) are provided as editing guidelines.

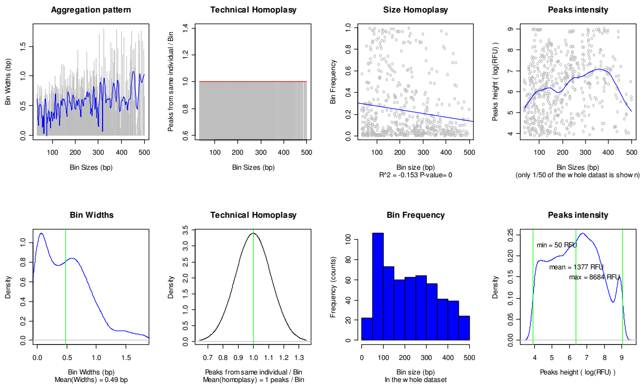


Fig 8. Quality statistics. The width, technical homoplasy and fluorescence of bins are displayed either according to the bin size (i.e. the position along the electropherogram) or as distributions. Statistics about size homoplasy are also provided.

5.2 Samples checking

RawGeno offers two displays for exploring scoring results (menu "RawGeno/Quality Check/Samples Checking"). The binary matrix can be directly visualized using a heatmap, showing individuals sorted according to their genetic relatedness. Alternatively, individuals can be examined with a principal coordinates analysis.

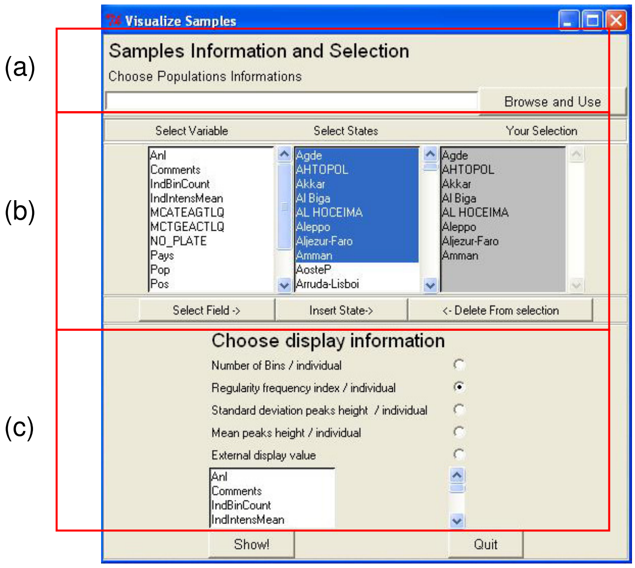
Both displays allow plotting quality statistics or external information (i.e. picked from the "info table" cited above) against the AFLP results. RawGeno computes four quality statistics regarding samples:

- The number of AFLP bands per sample,
- The outlier detection index, defined as the average frequency of the observed genotype.

$$\mathrm{R} = \frac{1}{\mathrm{nBin}} \sum_{\mathrm{i=1}}^{\mathrm{nBin}} P_i$$

Where nBin = number of AFLP bands in the sample, P_i = Frequency of the bin i in the whole dataset, that is present in the focal sample.

- The mean AFLP bands fluorescence per sample
- The standard deviation of AFLP bands fluorescence per sample



Samples Information and Selection

Choose Populations Informations

Select Variable: AnI, Comments, IndBinCount, IndIntensMean, MCATEAGTLQ, MCTGEACTLQ, NO_PLATE, Pays, Pop, Pos

Select States: Agde, AHTOPOL, Akkar, Al Biga, AL HOCEIMA, Aleppo, Aljezur-Faro, Amman, AosteP, Arruda-Lisboi

Your Selection: Agde, AHTOPOL, Akkar, Al Biga, AL HOCEIMA, Aleppo, Aljezur-Faro, Amman

Choose display information

Number of Bins / individual ○

Regularity frequency index / individual ●

Standard deviation peaks height / individual ○

Mean peaks height / individual ○

External display value ○

AnI, Comments, IndBinCount, IndIntensMean

Fig 9. Visualization device. (a) Selection of the info table (in order to compare external information with AFLP data), (b) selection and sorting of samples according to external information from the info table, (c) selection of quality statistics / external data to compare with the AFLP data. This device can produce either heatmaps of the AFLP binary matrix or principal coordinate analyses.

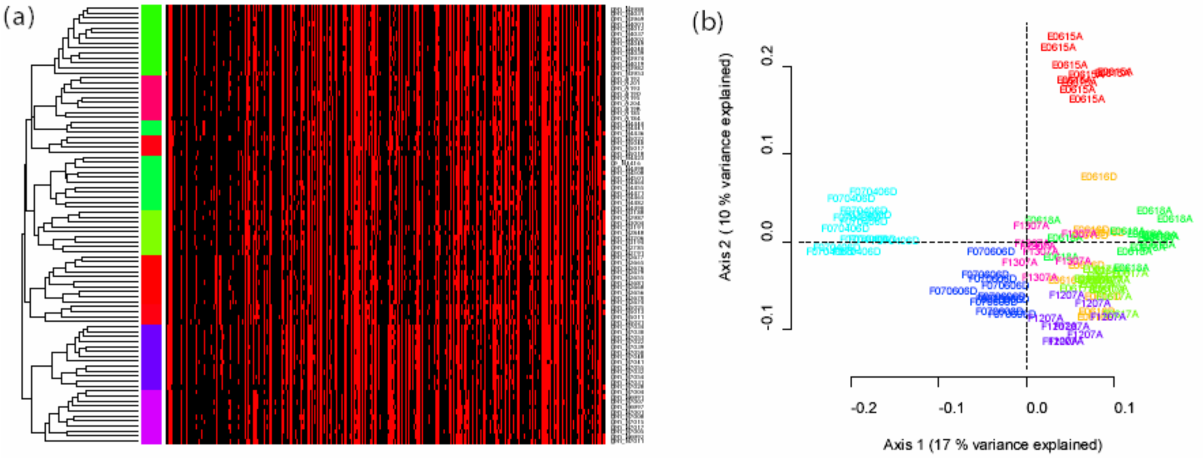
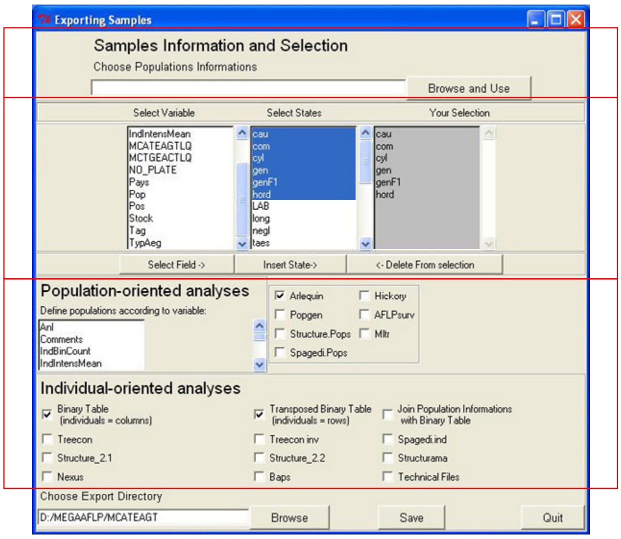


Fig 10. Visualizations of samples. RawGeno includes basic vizualisation devices for performing preliminary data mining. Specifically, results can be reviewed using (a) heatmaps of the binary matrix, where samples are sorted according to their genetic similarity and (b) principal coordinates analysis of the corresponding matrix. Both devices can compare AFLP results with either quality statistics (i.e. number of AFLP peaks per sample, mean and variance in fluorescence intensity and outlier detection index) or external information provided by users (e.g. the population from where samples were collected). In addition, both devices are handled through a graphical user interface for sorting and selecting samples to be visualized.

Note 5. Managing the dataset content. Users generally want to interactively manage their dataset along with visualizing results. Although the GUI gives some flexibility to this respect, the removal of outlier samples remains an operation that is not interactively available. Removing (or renaming) outlier samples within the info table provides a simple way for achieving this goal. RawGeno joins the info table with the scored AFLP dataset, therefore samples that are missing from the info table will not be displayed in visualization devices. Note that this "trick" only allows modifying results that are displayed as it does not affect the originally scored matrix nor the PeakScanner results. The production of the definitive dataset should start by removing undesirable samples before performing the PeakScanner analysis.

6 Exporting results

RawGeno produces exports in various formats *(8)*. These are accessible from the menu "4.Save / Export Currently Scored Datasets".



Exporting Samples

Samples Information and Selection

Choose Populations Informations

Select Variable	Select States	Your Selection
IndIntensMean MCATEAGTLQ MCTGEACTLQ NO_PLATE Pays Pop Pos Stock Tag TypAeg	cau com cyl gen genF1 hord LAB long negl taes	cau com cyl gen genF1 hord

Population-oriented analyses

Define populations according to variable: Anl, Comments, IndBinCount, IndIntensMean

☑ Arlequin ☐ Hickory
☐ Popgen ☐ AFLPsurv
☐ Structure.Pops ☐ Mltr
☐ Spagedi.Pops

Individual-oriented analyses

☑ Binary Table (individuals = columns) ☑ Transposed Binary Table (individuals = rows) ☐ Join Population Informations with Binary Table
☐ Treecon ☐ Treecon inv ☐ Spagedi.ind
☐ Structure_2.1 ☐ Structure_2.2 ☐ Structurama
☐ Nexus ☐ Baps ☐ Technical Files

Choose Export Directory: D:/MEGAAFLP/MCATEAGT

Adding sample information

Selecting samples to export

Selecting exporting formats

Fig 11. Results export device.

7 Merging data from several AFLP markers

The strategy is the following :

- Analyse each AFLP primer pair independently, and save it as a binary table (be consistent with the used data format, e.g. always save scored results as a binary table, with individuals as rows).
- Merge these binary tables into a single matrix using RawGeno

(menu Files / Import / Merge several datasets (Binary tables))

- Save this merged table using the usual saving GUI.

Note 6 All the datasets to be merged MUST have the same samples names. This is especially obvious when markers were multiplexed and therefore share the same *.fsa name, but this limitation must be checked carefully when merging data from various provenances.
Missing samples or different ordering of the scored dataset are allowed. Samples with missing genotypes are removed from the final dataset. This limitation can be bypassed by using the command line version:

i. Select files to merge

```
list.merge=choose.files(caption='Choose Files to Merge')
```

For linux users, use instead

```
list.merge=tk_choose.files(caption='Choose Files to Merge')
```

ii. Proceed to merging

```
MERGING(transpose = "indRows", exclude = T, replacewith = NA)
```

The transpose parameter states whether the binary matrices store individuals as lines ("indRows") or as columns ("indColumns"), the exclude parameter defines whether individuals that are not shared by all matrices will be removed from the final merged dataset (exclude = "T"). If kept (exclude = "F"), individuals with missing AFLP genotypes will be completed using NA values (replacewith = NA) when no data is available. Then proceed to export as explained above.

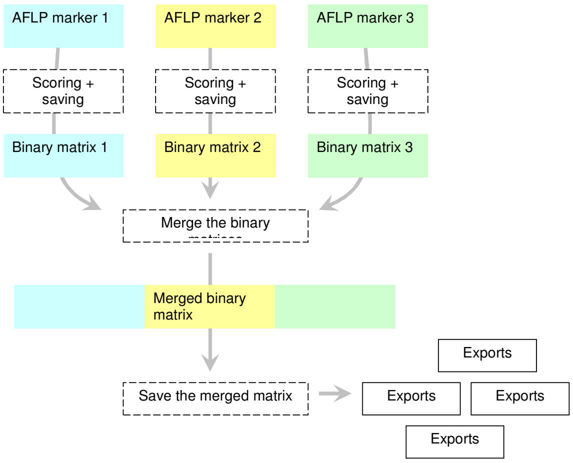


Fig 12. Merging several datasets. Strategy to merge several datasets using RawGeno. Note that this procedure requires to use the menu

"Files / Import / Merge several datasets (Binary tables)"

to proceed to merging, followed by the menu

"Save / Export Currently Scored Dataset"

to save results towards exports.

8 How to cite RawGeno?

Please cite RawGeno using the reference where it was first presented:

Arrigo N, Tuszynski JW, Ehrich D, Gerdes T, Alvarez N (2009) Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring. BMC Bioinformatics doi:10.1186/1471-2105-10-33.

9 Acknowledgements

We are grateful to Anahi Espindola, Matthias Borer, Anouk Beguin, Gregor Kozlowski, Aurélien Labhardt, Felix Gugerli, Sabine Brodbeck, Piya Kittipadakul, Julie B. Hébert, Julie Lee Yaw, Christian Parisod, Roland Dubillard and Robin Arnoux provided helpful comments during the development of RawGeno. This project was supported by the National Centre of Competence in Research (NCCR) Plant Survival and the PNR-59, two programs of the Swiss National Science Fundation.

10 References

1. Arrigo N, Tuszynski JW, Ehrich D, Gerdes T, Alvarez N (2009) Evaluating the impact of scoring parameters on the structure of intra-specific genetic variation using RawGeno, an R package for automating AFLP scoring. BMC Bioinformatics doi:10.1186/1471-2105-10-33.
2. Arrigo N, Felber F, Parisod C, Buerki S, Alvarez N, David J, Guadagnuolo R (in press) Origin and expansion of the allotetraploid *Aegilops geniculata*, a wild relative of wheat. New Phytol.
3. Whitlock R, Hipperson H, Mannarelli M, Butlin RK, Burke T (2008) An objective, rapid and reproducible method for scoring AFLP peak-height data that minimizes genotyping error. Mol Ecol Res 8:725-735.
4. Herrmann D, Poncet BN, 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. Genome 53(4):302-310.
5. Holland BR, Clarke AC, Meudt HM (2008) Optimizing automated AFLP scoring parameters to improve phylogenetic resolution. Syst Biol 57(3):347-366.
6. Vekemans X, Beauwens T, Lemaire M, Roldan-Ruiz I (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. Mol Ecol 11(1):139-151.
7. Paris M, Bonnes B, Ficetola GF, Poncet BN, Després L (2010) Amplified fragment length homoplasy: in silico analysis for model and non-model species. BMC Genomics doi:10.1186/1471-2164-11-287.
8. Ehrich D (2006) AFLPdat: a collection of R functions for convenient handling of AFLP data. Mol Ecol Notes 6: 603-604.