AFLP: generating objective and repeatable genetic data.

Thierry Onkelinx [1,*], **Pieter Verschelde**[1], **Sabrina Neyrinck**[1], **Gerrit Genouw**[1], **Paul Quataert**[1]

1. Research Institute for Nature and Forest (INBO), Scientific Services, Gaverstraat 4, B-9500 Geraardsbergen, Belgium
*Contact author: Thierry.Onkelinx@inbo.be

Keywords: Genetics, normalisation, classification, repeatability, randomisation

Amplified fragment length polymorphisms (AFLP) is a technique to get a DNA fingerprint from an organism. The total genomic DNA is split into fragments at fixed combinations of nucleotides, resulting in fragments of different lengths. After electropherese we get a DNA profile ("bands" or "peaks") of intensity versus fragment length. High intensity indicates the presence of fragment(s), low intensity the absence.

Our package AFLP, available on R-Forge[1], supports the analysis of AFLP data on several points: i) randomise the samples and add replicates, ii) normalise the intensity for differences due to laboratory handlings and equipment, iii) classify the normalised intensity into presence/absence data, iv) calculate the repeatability of the analysis and v) apply multivariate analysis on the classified data.

Replicates are added during the randomisation of the samples. This allows to measure the repeatability and helps the normalisation between batches. Therefore both within batch as between batch replicates are added. Bonin et al. (2004) suggest a "technical difference rate" (TDR) to estimate the repeatablity of two replicates from the same sample. We adapted this TDR formula to work with more than two replicates per sample. Furthermore we use this TDR not only to measure the repeatability of the samples, but also the fragments and the overall repeatability.

The normalisation is based on linear mixed models (**lme4**). The idea is to model the average intensity in function of the different sources of variability (DNA extraction, lab batch, capillary, ...). The raw residuals of the model are used as the normalised intensity. Several diagnostic plots are available to highlight potential problems. The normalised intensity is classified based on their distribution per fragment. The distribution is bimodal for a polymorph fragment (both presences and absences) and unimodal for a monomorph fragment (only presences or only absences). The lowest density between the modi is used as threshold for the classification. The user can get graphs displaying the distribution and the selected threshold. The build-in analysis capabilities are currently limited to methods for `hclust` and `princomp`. Both the normalised and classified data can be exported to other packages when needed.

Advantages of this package: i) The package starts from the design in the lab, ensuring the necessary replicates and randomisation. ii) The normalisation considers a much wider variety of sources than other software. iii) The intensity, typically lognormal distributed, can be transformed during the normalisation. iv) The classification threshold depends not an arbitrary threshold, but on the distribution of the normalised intensity. v) TDR is used as a measure of overall quality, quality per sample and quality per fragment.

References

Bonin, A., E. Bellemain, P. Bronken Eidessen, F. Pompanon, C. Borchmann, and P. Taberlet (2004). How to track and assess genotyping errors in population genetic studies. *Molecular Ecology 13*, 3261–3273.

[1] http://r-forge.r-project.org/projects/aflp/