Euphytica
DOI 10.1007/s10681-011-0368-3

Review and simulation of homoplasy and collision in AFLP

Gerrit Gort · Fred A. van Eeuwijk

Received: 7 July 2010 / Accepted: 21 January 2011

Abstract In this paper we give a short review of the problems of homoplasy and collision in AFLP, and describe a software tool that we developed to illustrate these problems. AFLP is a DNA fingerprinting technique, producing profiles of bands, the result of the separation of DNA fragments by length on a gel or microcapillary system. The profiles are usually interpreted as binary band absence/presence patterns. We focus on two major problems: (1) Within a profile two or more fragments of the same length but of different genomic origin may have been selected, colliding into a single band. This collision problem, akin to the birthday problem, may be surprisingly large. (2) In a pair of profiles two equally long fragments of different genomic origin may have been selected, appearing as identical bands in the two profiles. This is called homoplasy. Both problems are quantified by modeling AFLP as a random sampling technique of fragment lengths. AFLP may be used in phylogenetic studies to estimate the pairwise genetic similarity of individuals. Similarity coefficients like Dice and Jaccard coefficients overestimate the true genetic similarity because of homoplasy, with increasing bias for higher numbers of bands per profile. Corrected estimators are described, which do not suffer from bias. The ideas are illustrated using a new software tool. Data from studies on Arabidopsis and tomato serve as examples. Finally, we make some recommendations with respect to the use of AFLP.

Keywords AFLP · Collision · Homoplasy · Tomato · Rpanel · Similarity coefficient · Dice coefficient

G. Gort · F. A. van Eeuwijk
Biometris, Wageningen University and Research centre, P.O. Box 100, 6700 AC Wageningen, The Netherlands

G. Gort (✉)
Biometris , Wageningen University and Research Center, Radix building, room W2.Aa.063 Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands
e-mail: gerrit.gort@wur.nl

Introduction

AFLP is a commonly used DNA fingerprinting technique, developed by Vos et al. (1995). The name AFLP is interpreted as an acronym of Amplified Fragment Length Polymorphism, giving an indication of it's working: it aims to find polymorphisms in lengths of selected DNA fragments, which are amplified by PCR. AFLP is used in many fields of the life sciences, but the majority of applications are found in the Plant Sciences. In Fig. 1 an example of an AFLP gel is shown, originating from a project on tomato quality within the Dutch Center for BioSystems Genomics. The columns (lanes) of the gel contain DNA fragments of different tomato genotypes. Bands within the lanes represent DNA fragments of specific lengths, shorter fragments further down.

Published online: 03 February 2011

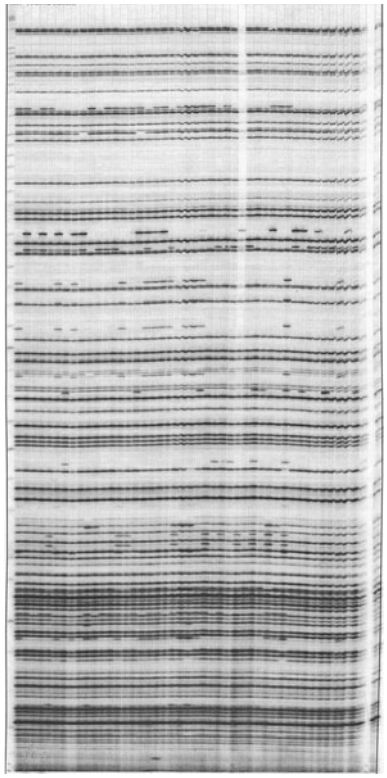

Fig. 1 Example of an AFLP gel, originating from a project on tomato quality within the Dutch Center for BioSystems Genomics, created by Keygene NV; the first column (*lane*) contains a size ladder of DNA fragments with known fragment lengths; columns 2–48 contain bands (DNA fragments) of 47 different tomato genotypes

Originating in the early 90s of the 20th century, in the dynamic era of genetics and bioinformatics, AFLP may be considered quite old. The title of the review paper by Meudt and Clarke (2007) "Almost Forgotten or Latest Practice?" suggests the same. A simple way to check the present scope of AFLP is to count the number of publications, making mention of it. Figure 2 shows the yearly number of scientific papers referring to the AFLP procedure. The figure demonstrates that the application of AFLP, after a quick rise around the change of the century, currently remains at a constant, high level.

Having sketched the place and scope of AFLP in the field of the life sciences, we introduce the topic of this paper. The result of AFLP is a profile of bands, like a bar code, in different lanes of an electrophoretic gel or microcapillary system. Usually, the band information is interpreted binary, i.e. as band absent/

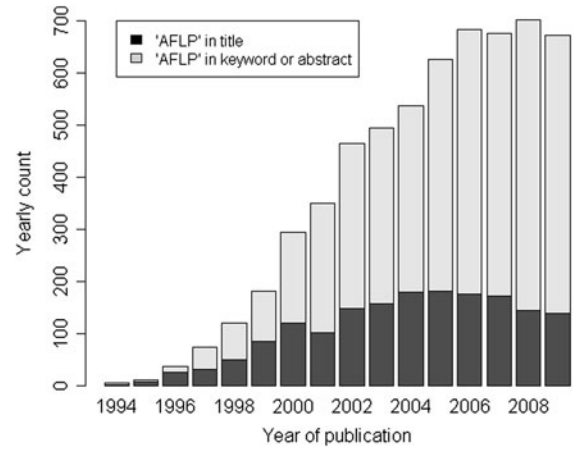

Fig. 2 Yearly counts of scientific publications in Web of Science in the period 1994–2009, containing the phrase "AFLP" in title, keyword or abstract

present patterns on discrete positions within the lane. A band is supposed to represent a unique DNA fragment. Corresponding bands in different lanes are supposed to be homologous, that is, the DNA fragments are identical and originate from the same genomic locus. The DNA fragments, however, are largely anonymous: only the ends of the fragments, to which the primers bind, have known nucleotide sequences, but the interior nucleotides of the fragments remain unknown. It could be that within a lane (i.e. a single genotype) a band, supposedly representing a single DNA fragment, may actually correspond to *multiple* fragments, because equally sized, but different fragments were selected, which comigrated to the same position within the lane. We call this *collision*. If two lanes (i.e. two genotypes) are compared, two bands, supposedly representing two identical fragments from the two genotypes, may actually represent two *different* fragments, because two different but equally sized fragments were selected, and comigrated to the same position in the two lanes. This problem is called *homoplasy*. In the literature, the type of comigration that we call collision, is sometimes called masking (Meudt and Clarke 2007), but more often is also referred to as homoplasy, causing confusion about the actual topic of study. Therefore, we find it useful to distinguish between the problems of comigrating fragments within a lane and between two lanes, and name them collision and homoplasy, respectively.

Homoplasy has been pinpointed as a major issue in the interpretation of AFLP data (Bonin et al. 2007; Meudt and Clarke 2007). Various ways of assessing homoplasy (or collisions) can be found in the literature:

1. In-silico AFLP and Monte Carlo simulation. For species with known sequences in-silico AFLP can be performed, mimicking the AFLP procedure on the computer. Examples of in-silico AFLP studies are Althoff et al. (2007), Stölting et al (2009), Gort et al. (2009), Caballero and Quesada (2010), and Paris et al. (2010). In Monte Carlo simulation studies, AFLP is simulated by sampling from a given fragment length distribution. Examples of Monte Carlo simulation studies are Vekemans et al. (2002), Koopman and Gort (2004).
2. Single nucleotide primer extension. To identify the anonymous interior nucleotides, an AFLP primer is elongated with each of the four single nucleotides A, C, T, and G, and new AFLP profiles are made. This makes it possible to identify the next nucleotide(s), thereby recognizing multiple fragments within a band or homoplasious fragments between bands. Studies of this type include Hansen et al. (1999) and O'Hanlon and Peakall (2000).
3. Sequencing of fragments. AFLP bands are cut out of the gel, re-amplified and cloned into bacteria. Bacterial plasmid DNA is sequenced, resulting in the nucleotide sequence of the captured AFLP fragments. Different clones (bacterial colonies) per band are used. Sometimes only a few clones are taken, making it doubtful whether all different fragments are sequenced in case of collision. Studies of this type include Rouppe van der Voort et al. (1997), Mechanda et al. (2004) and Ipek et al. (2006).

The general conclusions we draw from these studies are: (1) collision occurs regularly, with larger rates for profiles with more bands; (2) homoplasy occurs regularly, with larger rates for more distantly related individuals. Both collision and homoplasy are probably underreported due to insufficient sequencing efforts.

All the described studies are case studies on specific organisms, or sets of organisms, which in our view lack generality. We therefore present a different approach: by modeling AFLP from a statistical point of view, we are able to *estimate* the level of collision and homoplasy. This brings as benefits the allowance to (1) predict the level of collision and homoplasy in *any* case; (2) correct derived quantities, like similarity coefficients, for homoplasy and collision.

We summarize our earlier findings on homoplasy and collision in AFLP, as published elsewhere. We furthermore describe a new visualization tool for illustration of collision and homoplasy in AFLP, which may help researchers to judge the extent and seriousness of the problems.

Materials and methods

AFLP modeled

Descriptions of the AFLP procedure can be found in many papers (e.g. Mueller and LaReesa Wolfenbarger 1999; Vuylsteke 2007). Here we describe the main three steps of AFLP (DNA digestion, fragment selection, and fragment separation/visualization) from a statistical point of view:

(1) DNA is digested by two restriction enzymes, and adaptors are ligated to the resulting fragments. The relative frequencies of the lengths of these sequences form a probability distribution, which is heavily asymmetrical (see next section). Fragments with lengths between a given lower bound (≈ 50) and upper bound (≈ 500) and at least one labeled restriction site, are eligible for selection and visualization. We call this subset of fragments the *population of candidate fragments*. This population consists of many, many thousands of elements. The population size is roughly proportional to the genome size. The lengths of the fragments form the *fragment length distribution* (fld).

(2) Primers make a selection of fragments for amplification by PCR. This selection process can be considered as *systematic sampling* of fragments from the population of candidate fragments. With respect to fragment lengths, we consider it as *random sampling* of lengths from the fld. Given a set of primers, the sample size is proportional to the size of the population of candidate fragments. Usually the primers are chosen such that the number of fragments per lane is between 50 and 150. Because this sample size is small compared to the population size, we consider it as sampling with replacement.

(3) The sampled fragments are separated by length using electrophoresis and visualized as bands. We assume that the position of a fragment within a profile

is largely determined by its length, so that two fragments of equal length will show up as a single band. The end product is a vector of binary data: bands at discrete positions within a lane are either absent or present. The presence of a band at position *i,* indicates that *at least* one fragment of corresponding length was selected.

Two related genotypes share parts of their DNA. Therefore, they share part of their two populations of candidate fragments (with sizes M_1 and M_2), formed after step 1 of the AFLP procedure. We call this common part the population of common fragments (with size M_c). The sets of fragments unique to each genotype are called the populations of unique fragments (with sizes M_{u1} and M_{u2}). The pairwise genetic similarity is then defined as the weighted average of fractions of common fragments: $w_1\frac{M_c}{M_1} + w_2\frac{M_c}{M_2}$ with the relative population sizes as weights ($w_1 = {}^{M_1}/_{(M_1+M_2)}$ and $w_2 = 1 - w_1$). This definition corresponds to the Dice coefficient of similarity. For a more elaborate description see Gort et al. (2009).

Fragment length distribution

The lengths of the fragments in the population of candidate fragments form the fragment length distribution (fld). The fld is highly asymmetric: shorter fragments are much more likely than longer ones. The fld can be estimated in different ways. Innan et al. (1999) describe a method based upon random nucleotide order to arrive at a truncated geometric distribution. Koopman and Gort (2004) estimate the fld based on in-silico AFLP, searching for restriction sites in the complete genome of *Arabidopsis thaliana*. The resulting distribution should be reasonable for genomes with a GC content close to 36%. It is highly asymmetrical, with the shortest fragment more than 20 times more likely than the longest (which is 450 nucleotides longer). Fld's of other species using in-silico AFLP were studied by Caballero and Quesada (2010). A third estimation method of fld's, as described by Gort et al. (2006), is based directly on the binary AFLP data within a profile. In this method, the binary band absence/presence data are modeled as a smooth, monotone decreasing function of the fragment lengths, using a generalized linear model (McCullagh and Nelder 1991) and monotone P-splines (Bollaerts et al. 2006).

Collision

In step 2 of the AFLP procedure two or more fragments of the same length may have been drawn from the fld, which collide in a single AFLP band. This problem is akin the birthday problem. The birthday problem asks how many persons are needed in a group, to have a probability of at least ½ for two or more persons to share a birthday. This number is remarkably low: 23. In AFLPs the situation is worse due to the skewed fld: in a typical AFLP situation only 19 fragments are needed to have a probability $>$ ½ for at least one collision to occur. In Gort et al. (2006, 2008, 2009) it is explained how the total number of collisions in a profile and the probability of a band to contain a collision can be estimated, given the number of fragments, the number of bands, or the band positions in a profile. The total number of collisions mainly depends on the sample size of fragments within a profile. In a typical AFLP situation with 80 bands it is in the range 16 ± 5. Collisions tend to concentrate at the smaller fragment lengths. In a typical AFLP case the probability of a band to contain a collision is for the shortest band position more than 20 times larger than for the longest.

Homoplasy

If AFLP profiles of two related genotypes are compared, some bands will be common, whereas others are unique to the genotypes. Part of the common bands, however, are common *due to chance*: two different but equally sized fragments were sampled from the two populations of unique fragments, resulting in homoplasious bands. Gort et al. (2008, 2009) describe how the number of common (homologous) fragments may be estimated. The trick is to compare the sum of the estimated total numbers of fragments in the two profiles separately with the estimated number of fragments in the profile, obtained by overlaying the two separate profiles. The difference is an estimate of the number of common fragments in the sample. Estimates of the numbers of unique fragments are obtained simultaneously. It is obvious that closely related genotypes will not suffer heavily from homoplasy, because fragments drawn from the population of common fragments dominate, and the probability that two equally sized fragments from the two small populations of unique fragments were sampled

simultaneously is small. At the other extreme, for unrelated genotypes fragments drawn from the unique parts dominate, and the probability of homoplasious fragments is much larger.

Corrected similarity coefficients

Homoplasy causes similarity coefficients like Dice or Jaccard, based on binary band data, to overestimate the true similarity, resulting in positive bias. This bias depends on the numbers of fragments in the profiles, and is larger for less related genotypes. Once estimates of the numbers of common and unique fragments in the sample are obtained, as described in the previous section, modified similarity coefficients may be calculated, replacing the band counts by estimated fragment counts (Gort et al. 2009). The resulting coefficients are unbiased, and have smaller standard errors for most (but not all) cases.

Interactive visualization of collision and homoplasy

We made two programs to visualize in an interactive way the results on collision and homoplasy, using the rpanel package (Bowman 2007) of the R program (R Development Core Team 2005). These programs are available from the authors. For both programs we used the fld, estimated from *Arabidopsis thaliana* using in-silico AFLP (see Sect. Fragment length distribution) with scoring range 51–500. If the interest is in an organism with GC content substantially deviating from the 36% of *A. thaliana*, a different fld should be used. This would require estimation of an appropriate fld, and replacing the present fld by the new one.

In the first program, labeled AFLPcollision.r, the random sampling of DNA fragments from a fragment length distribution is simulated. An example is shown in Fig. 3. The sampled fragments are sorted on length within the lane and visualized. We use a square root scale for fragment lengths within a lane, which is close to what is observed in practice. Sampling of fragments can be done fragment by fragment, or in groups of five or ten. The occurrence of a collision is notified by coloring the band within the lane in red, with darker colors indicating higher order collisions. The number of fragments of a specific length are also kept track of and shown besides the band outside the lane. Optionally, the annotated version of the visualized AFLP profile can be replaced by a not-annotated one, which is comparable to an AFLP profile in practice. The total number of selected fragments, the total number of bands, and total number of collisions are shown. Optionally the number of fragments and number of collisions can be estimated using the methodology as described in Gort et al. (2009), for comparison with the true number in the simulated AFLP profile. The influence of the fld can be studied by comparison of collision results of the highly asymmetric *A. thaliana* distribution with those of the uniform distribution. The uniform distribution (which gives all fragment lengths equal probability) corresponds to a best case scenario, resulting in a lower bound for the expected number of collisions and homoplasious bands. Any other fld will result in more collisions and homoplasious bands on average.

In the second program (AFLPhomoplasy.r) we randomly sample fragments from a fragment length distribution for a pair of AFLP profiles, again allowing fragments to be sampled individually, or in groups of five or ten. An example is shown in Fig. 4. Each fragment is either sampled from the common part of the two populations of candidate fragments, in which case a band will appear in both lanes, or from one of the two unique parts, resulting in a band in only one of the two lanes. The fraction of common fragments can be specified using a slider. The size ratio of the populations of candidate fragments can be chosen, ranging from 1 (equally sized genomes) to 10 (first genome ten times larger than second). From the fraction common fragments and relative population sizes, the pairwise genetic Dice similarity is calculated. Any fragment drawn from the common part will result in two green bands at the same position in the two lanes. A fragment drawn from one of the unique parts will appear as a black band in the respective lane. Collisions are shown as before, using a circle with the count of equally sized fragments besides the lane. If two homoplasious fragments are sampled (hence originating from the two unique parts of the fragment populations), the resulting bands in the two lanes are colored red. Optionally a not-annotated version of the AFLP profile can be shown. All relevant information is kept track of, like fragment, band, and collision count per lane, true number of common fragments, observed number of common bands and true number of homoplasious bands. The ordinary Dice coefficient

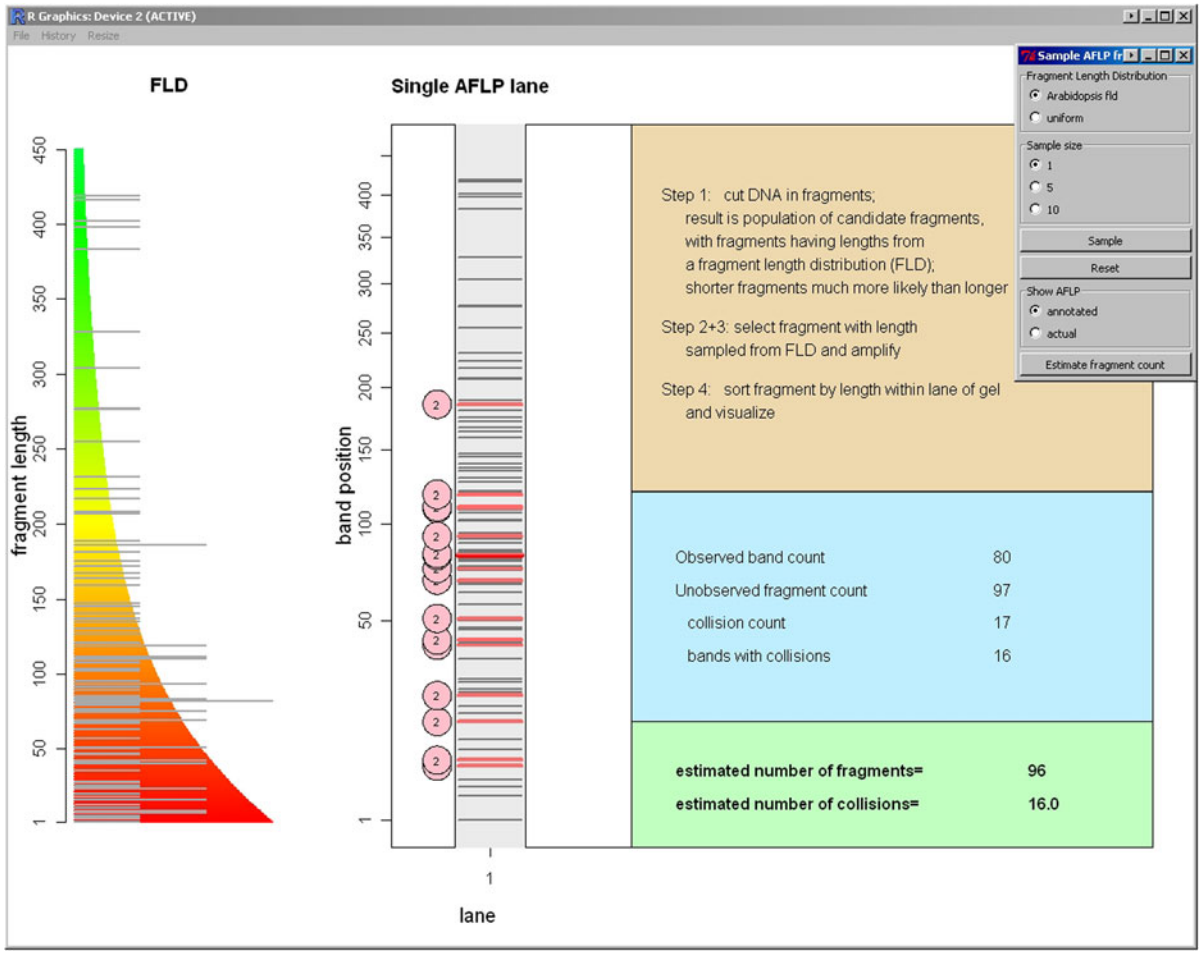

Fig. 3 Screenshot of the AFLPcollision program for a typical AFLP profile with 80 bands; the *left* part shows the fld (estimated from *A. thaliana*) with counts of sampled fragment lengths shown as bars; the middle part shows the annotated AFLP profile, with *black colored bands* indicating band without collision, *red colored bands* indicating collisions, and collisions counts shown in *circles* to the *left*; the *right* part describes the AFLP sampling, shows statistics for the current profile, and estimates fragment counts and collision counts according to the methods described in the paper; the *top right* window allows the setting of parameters, and the performance of actions, like sampling fragments; this window can be moved anywhere on the screen

based upon band counts is shown. Optionally the corrected Dice coefficient is calculated, as described in Gort et al. (2009).

Example applications of the programs

We have simulated AFLP profiles for a number of scenario's. In most cases fragments were sampled until the observed number of bands per lane was close to 80, using the *A. thaliana* fld. We call this a "typical" AFLP profile. The lanes of the AFLP gel on tomato, shown in Fig. 1, are in this respect quite typical. A single application of the AFLPcollision.r program is shown, and three applications of AFLPhomoplasy.r, for three distinct biologically relevant cases.

Figure 3 shows an application of program AFLPcollision.r In this example we sampled 97 fragments to arrive at 80 bands in the lane. Therefore, 17 fragments had lengths already sampled earlier (17 collisions). The first collision was obtained at the 15th sampled fragment (not shown). In the final profile one fragment length occurred three times. We

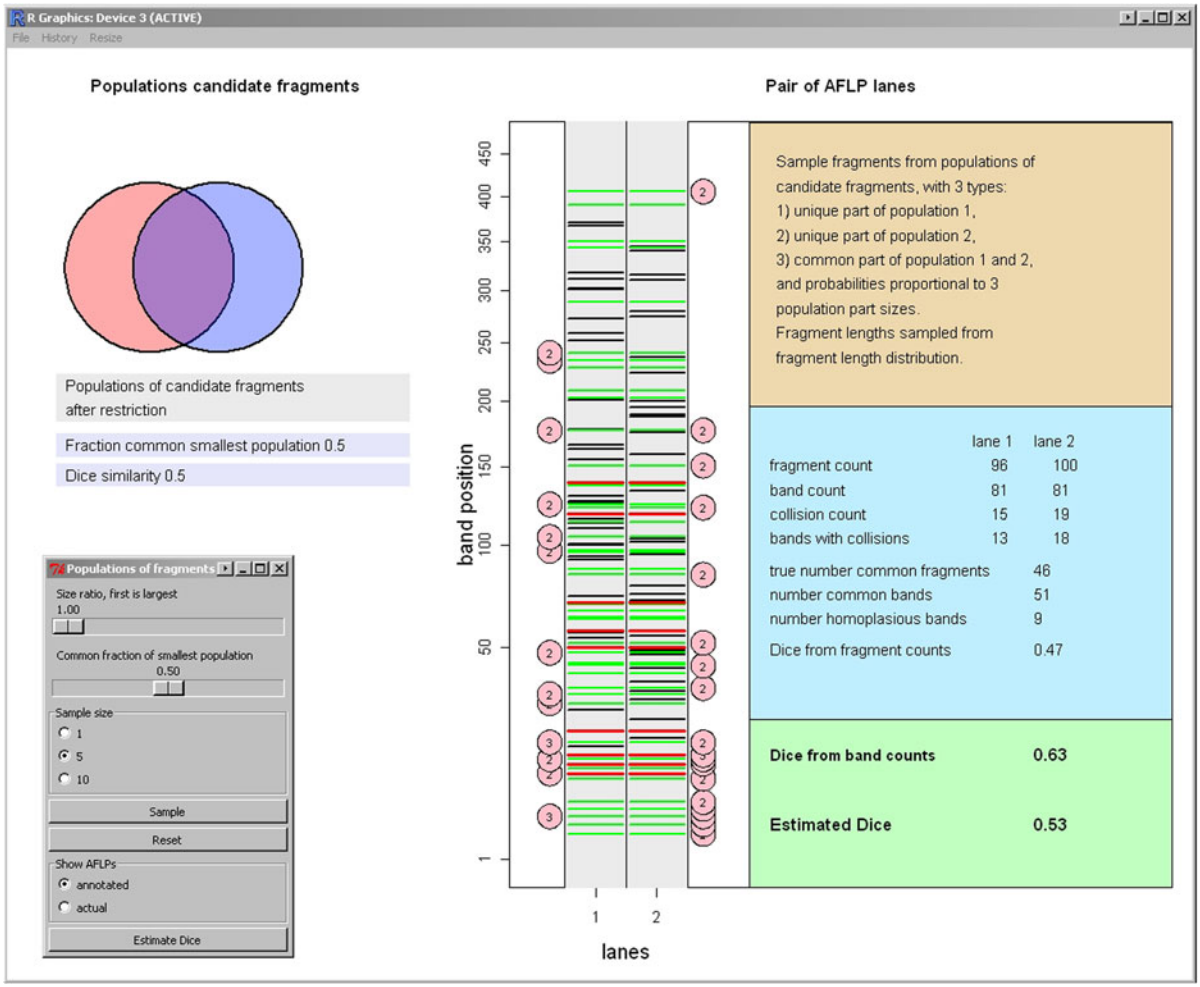

Fig. 4 Screenshot of the AFLPhomoplasy program for two related genotypes, equally sized genomes, fraction common fragments 0.5; the *left* part shows the two populations of candidate fragments as partly overlapping *circles*; the *middle* part shows the two annotated AFLP profiles with collision counts in *circles* besides the profiles; the coloration of bands now indicates whether fragments were homologous (*green*), homoplasious (*red*), or unique to the genotypes (*black*); the right part again describes the AFLP sampling, shows statistics for the current two profiles, and calculates the Dice coefficients based on bands and estimated fragment counts according to the methods described in the paper; the bottom left window allows the setting of parameters, and the performance of actions, like sampling fragments

call the occurrence of more than two fragments of equal length a higher order collision. In total 16 bands contained two or more fragments. The estimated number of fragments based on the band positions was 96.0, slightly lower than the true number (=97).

In Fig. 4 an application of program AFLPhomoplasy.r is shown. We chose in this example the fraction of common fragments to be equal to 0.5 and equally sized genomes, so that half of the fragments in the populations of candidate fragments (after digestion of the DNA in step 1 of the AFLP procedure) were chosen to be identical. This situation is indicated in the screenshot at the top left-hand side as two equally sized circles overlapping for 50%. In the first lane 96 fragments, and in the second lane 100 fragments were drawn, resulting in 81 bands in both lanes. In the two lanes 15 and 19 collisions occurred, with 13 and 18 bands containing collisions, so that again some higher order collisions took place. In total 46 fragments were truly common (green bands), which would result in a Dice coefficient of 0.47 (had the fragments directly been observed), close to the population value 0.5. The

total number of common bands, however, was 51, of which nine were homoplasious (red bands). The Dice coefficient calculated from the band information was 0.63, much higher than the true value 0.5. The corrected Dice coefficient was 0.53, still slightly larger than the population value.

Figure 5 again shows an application of AFLPhomoplasy.r for a pair of genotypes, with equally sized genomes, but now with a high fraction of common fragments (0.85), as indicated by the largely overlapping circles. Such value could be plausible in a typical AFLP study of related genotypes, like the different tomato cultivars shown in Fig. 1. In the two lanes of this example 96 and 99 fragments were drawn, resulting in 79 and 82 bands, both with 17 collisions. In both lanes 15 bands contained collisions, so that again higher order collisions occurred, as could also be seen in the annotation in pink besides the bands. The set of truly common fragments (green bands) numbered 82, which would result in a Dice coefficient of 0.84, close to the population value 0.85. The total number of observed common bands was 70, of which only two were homoplasious (red bands). The Dice coefficient calculated from the band information was 0.87, slightly larger than the population value 0.85. The corrected Dice coefficient was 0.84. In this example with closely related genotypes, homoplasy is a minor problem. We note that in practice researchers often base the similarity coefficient on the non-monomorphic bands in a collection of AFLP profiles only, resulting in a lower Dice coefficient than the ones reported here.

In Fig. 6 an application of AFLPhomoplasy.r is given for the case of equally sized genomes, but now with the fraction of common fragments equal to 0 (circles do not overlap at all). In this instance all fragments from the two populations of candidate fragments are supposed to be different. This situation, which is taken as starting point in Koopman and Gort (2004), is rather unrealistic, but is interesting because it shows a worst case scenario: *all* observed similarity is due to chance alone. In the output from the program no green bands can be found anymore, but black bands (representing bands unique for a genotype) and red bands (for homoplasious bands) only. In the example 98 and 90 fragments were sampled to arrive at 82 and 78 bands. Collision counts were 16 and 12 respectively, with again higher order collisions (even a four-fold collision). All 24 common bands were necessarily homoplasious. The extremely biased Dice coefficient based on band information is 0.3. The corrected Dice coefficient is 0.023, close to the true value 0.

Conclusions and discussion

In this paper we have summarized and illustrated results on two major problems in the interpretation of AFLP data: collision and homoplasy. We modeled AFLP as a random sampling procedure of fragment lengths from a fld. Given the fld, the total number of collisions in a profile can be estimated. This number can be surprisingly high, depending on the total number of bands per lane. In a typical AFLP profile with 80 bands 16 (± 5) collisions may have occurred. Probabilities of a band to contain a collision can be calculated. Collisions tend to concentrate at the smaller fragment lengths. In a typical AFLP profile the shortest band is ≈ 20 times more likely to contain a collision than the longest.

In profiles of two related genotypes homoplasy may occur, depending on the relatedness of the genotypes and numbers of bands in the lanes. Less relatedness and more bands per lane result in more homoplasious bands. The number of truly common fragments in the pair of profiles can be estimated.

Similarity coefficients based upon the band counts, estimating the genetic similarity between genotypes, are biased estimators: they overestimate the true similarity. Corrected similarity coefficients, using the estimated numbers of common fragments, are formulated, which correct the bias due to homoplasy and collision.

The software tool, presented in this paper, is developed using the rpanel package of the statistical program R. It illustrates the above mentioned problems in AFLP, treating it as random sampling of fragment lengths from a fld. A number of scenario's have been worked out as examples, showing some typical outcomes.

We can see different aspects of relevance of our work:

1. The software tool may be helpful to appliers of AFLP to become aware of the potential size of the problems of homoplasy and collision in their AFLP profiles. Recognition of the size of the

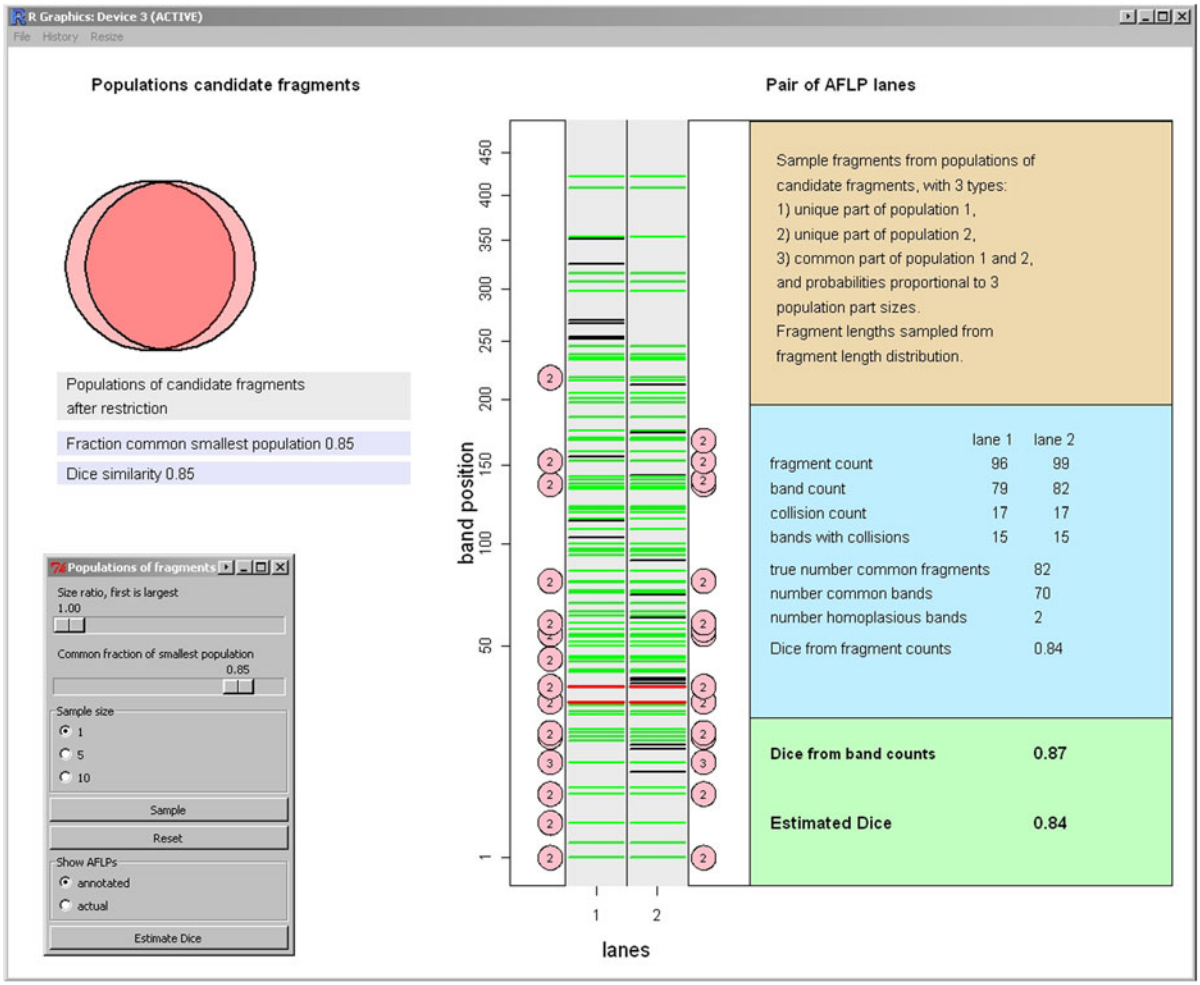

Fig. 5 Screenshot of the AFLPhomoplasy program for two related genotypes, equally sized genomes, high fraction common fragments 0.85

problem will lead to better understanding of the data and its potentially unexpected characteristics.

2. Refinements in the design of AFLP studies are suggested. If a genotypic interpretation of bands is important, like in QTL studies, it may be better to use highly selective primers, limiting the number of bands per lane, and thereby limiting problems. In that case the advise is to go for quality, not for quantity. Our results also allow the applier to pinpoint possibly problematic bands. A set of recommendation along these lines can be found in Gort et al. (2008) and Paris et al. (2010).
3. By modeling AFLP in a general way, we can quantify the extent of collision and homoplasy, not targeting any special case. Therefore, we are able to suggest corrections for derived quantities, like the corrected similarity coefficients described earlier.
4. Our work widens the applicability of AFLP. In the past, the general advise was to use AFLP only for studies of closely related taxa, as citations from Althoff et al. (2007) ("AFLP data are best suited for examining phylogeographic patterns within species and among very recently diverged species") and O'Hanlon and Peakall (2000) ("Studies of phylogeny with AFLPs are therefore only suited to closely related taxa.") show. However, the problems of collision and homoplasy will always occur, with a smooth transition from small problems in case of AFLP profiles with few bands

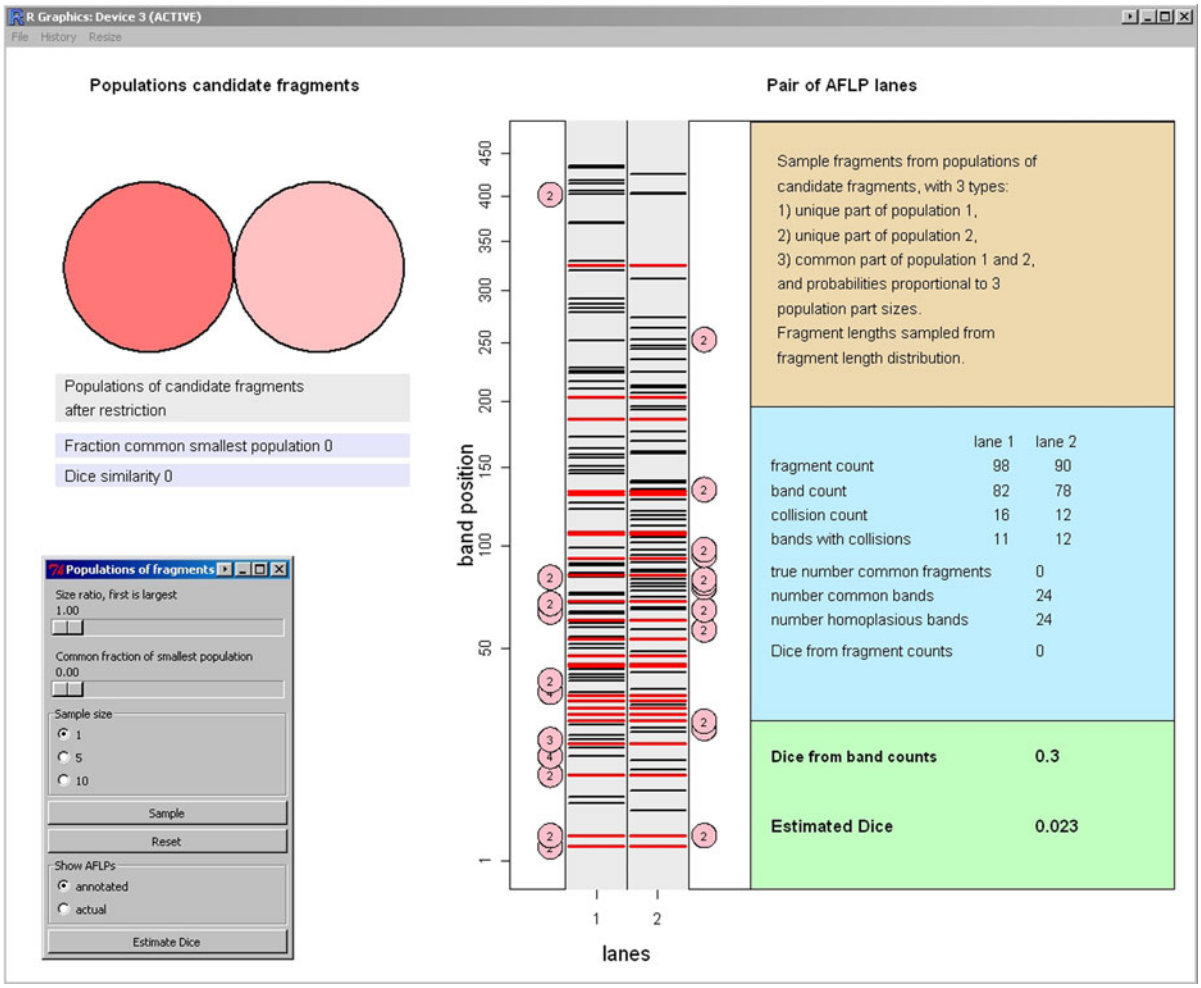

Fig. 6 Screenshot of the AFLPhomoplasy program for two unrelated genotypes (fraction of common fragments 0), equally sized genomes

and closely related taxa, to large problems in case of profiles with many bands and distantly related taxa. The rather artificial dichotomy into situations appropriate for AFLP studies, pretending that problems are non-existing, and inappropriate situations for AFLP studies is suboptimal. Corrections for homoplasy and collisions allow AFLP to be used in a wider range of studies with more reasonable results. This becomes extra relevant at present, where association studies are performed using association panels, consisting of diverse collections of genotypes with little knowledge about their genetic relationships.

Point of concern in our treatment of AFLP is the assumption that equally sized fragments travel equally far in a gel or microcapillary system, and hence end up at the same position within the lane(s). The electrophoretic separation of fragments is indeed mainly, but not solely, by size. From empirical studies it appears that slightly shorter or longer fragments may travel the same distance. However, different studies do not produce univocal conclusions: Meksem et al. (2001) report that all fragments per band were equally sized, Ipek et al. (2006) find rather small differences in lengths, but Mechanda et al. (2004) report huge differences in lengths. More study is needed here. It is unclear how this will influence our results. We could argue that results will remain approximately the same, because some equally long fragments may arrive at different distances, but some shorter or longer

fragments will arrive instead. And hence, the net effect may be roughly nil. Our assumption that fragments arrive at discrete distances corresponding to basepair lengths within a lane is also questionable. Maybe with a better scoring algorithm with sub-basepair resolution (Holland et al. 2008), part of the homoplasy could be prevented from the start.

An interesting connection with collisions can be made if AFLP data are scored *codominantly*. So far, AFLP profiles were *dominantly* scored, i.e. interpreted binary as band absence/presence patterns. A band simply indicated the presence of a DNA fragment of specific length, irrespective of the copy number: the fragment could have occurred with two copies (homozygous) or one copy (heterozygous), in the diploid case. However, AFLP bands may be scored *codominantly*, interpreting the band intensity. A band with higher intensity indicates more amplified DNA, plausibly explained by two copies of the fragment (homozygous). The copy number of a fragment may be inferred by fitting normal mixture models (see e.g. Piepho and Koch 2000; Gort and van Eeuwijk 2010). However, collisions yield more amplified DNA as well, and hence produce darker bands. Paris et al. (2010) report in an empirical study that bands with collisions tended to exhibit higher intensities than bands without, but the variation in intensity was very high. Therefore, it will be challenging to disentangle the effects of copy number and collision in the codominant interpretation of AFLP data.

Acknowledgements

We would like to thank Paul Eilers, for pointing us to the rpanel package, and for his encouraging enthusiasm in general.

References

Althoff DM, Gitzendanner MA, Segraves KA (2007) The Utility of amplified fragment length polymorphisms in phylogenetics: a comparison of homology within and between genomes. Syst Biol 56:477–484

Bollaerts K, Eilers PHC, van Mechelen I (2006) Simple and multiple P-splines regression with shape constraints. Br J Math Stat Psychol 59:451–469

Bonin A, Ehrich D, Manel S (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists. Mol Ecol 16:3737–3758

Bowman A, Crawford E, Alexander G, Bowman RW (2007) rpanel: simple interactive controls for R functions using the tcltk package. J Stat Softw 17(9)

Caballero A, Quesada H (2010) Homoplasy and distribution of AFLP fragments: an analysis in silico of the genome of different species. Mol Biol Evol 27:1139–1151

Gort G, van Eeuwijk FA (2010) Codominant scoring of AFLP in association panels. Theor Appl Genet 121:337–351

Gort G, Koopman WJM, Stein A (2006) Fragment length distributions and collision probabilities for AFLP markers. Biometrics 62:1107–1115

Gort G, Koopman WJM, Stein A, van Eeuwijk FA (2008) Collision probabilities for AFLP bands, with an application to simple measures of genetic similarity. JABES 13:177–198

Gort G, van Hintum T, van Eeuwijk F (2009) Homoplasy corrected estimation of genetic similarity from AFLP bands, and the effect of the number of bands on the precision of estimation. Theor Appl Genet 119:397–416

Hansen M, Kraft T, Christiansson M, Nilsson N-O (1999) Evaluation of AFLP in *Beta*. Theor Appl Genet 98: 845–852

Holland BR, Clarke AC, Meudt HM (2008) Optimizing automated AFLP scoring parameters to improve phylogenetic resolution. Syst Biol 57:347–366

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

Ipek M, Ipek A, Simon PW (2006) Sequence homology of polymorphic AFLP markers in garlic (*Allium sativum* L). Genome 49:1246–1255

Koopman WJM, Gort G (2004) Significance tests and weighted values for AFLP similarities, based on arabidopsis in silico AFLP fragment length distributions. Genetics 167: 1915–1928

McCullagh P, Nelder JA (1991) Generalized linear models, 2nd edn. Chapman & Hall, London

Mechanda SM, Baum BR, Johnson DA, Arnason JT (2004) Sequence assessment of comigrating AFLP (TM) bands in Echinacea—implications for comparative biological studies. Genome 47:15–25

Meksem K, Ruben E, Hyten D, Triwitayakorn K, Lightfoot DA (2001) Conversion of AFLP bands into high-throughput DNA markers. Mol Genet Genomics 265:207–214

Meudt HM, Clarke AC (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. Trends Plant Sci 12:106–117

Mueller UG, LaReesa Wolfenbarger L (1999) AFLP genotyping and fingerprinting. Trends Ecol Evol 14:389–394

O'Hanlon PC, Peakall R (2000) A simple method for the detection of size homoplasy among amplified fragment length polymorphism fragments. Mol Ecol 9:815–816

Paris M, Bonnes B, Ficetola GF, Poncet BN, Despres L (2010) Amplified fragment length homoplasy: in silico analysis for model and non-model species. BMC Genomics 11:13

Piepho HP, Koch G (2000) Codominant analysis of banding data from a dominant marker system by normal mixtures. Genetics 155:1459–1468

R Development Core Team (2005) R: a language and environment for statistical computing. R foundation for statistical computing, http://www.R-project.org, Vienna, Austria

Rouppe van der Voort JNAM, van Zandvoort P, van Eck HJ, Folkertsma RT, Hutten RCB, Draaistra J, Gommers FJ, Jacobsen E, Helder J, Bakker J (1997) Use of allele specificity of comigrating AFLP markers to align genetic maps from different potato genotypes. Mol Gen Genet 255:438–447

Stölting KN, Gort G, Wust C, Wilson AB (2009) Eukaryotic transcriptomics in silico: optimizing cDNA-AFLP efficiency. BMC Genomics 10:565

Vekemans X, Beauwens T, Lemaire M, Roldán-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 11:139–151

Vos P, Hogers R, Bleeker M, Reijans M, Vandelee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Vuylsteke M (2007) AFLP technology for DNA fingerprinting. Nat Protoc 2:1387–1398