Theor Appl Genet (2009) 119:397–416
DOI 10.1007/s00122-009-1047-9

ORIGINAL PAPER

Homoplasy corrected estimation of genetic similarity from AFLP bands, and the effect of the number of bands on the precision of estimation

Gerrit Gort · Theo van Hintum · Fred van Eeuwijk

Received: 2 July 2008 / Accepted: 21 April 2009 / Published online: 13 May 2009

Abstract AFLP is a DNA fingerprinting technique, resulting in binary band presence–absence patterns, called profiles, with known or unknown band positions. We model AFLP as a sampling procedure of fragments, with lengths sampled from a distribution. Bands represent fragments of specific lengths. We focus on estimation of pairwise genetic similarity, defined as average fraction of common fragments, by AFLP. Usual estimators are Dice (D) or Jaccard coefficients. D overestimates genetic similarity, since identical bands in profile pairs may correspond to different fragments (homoplasy). Another complicating factor is the occurrence of different fragments of equal length within a profile, appearing as a single band, which we call collision. The bias of D increases with larger numbers of bands, and lower genetic similarity. We propose two homoplasy- and collision-corrected estimators of genetic similarity. The first is a modification of D, replacing band counts by estimated fragment counts. The second is a maximum likelihood estimator, only applicable if band positions are available. Properties of the estimators are studied by simulation. Standard errors and confidence intervals for the first are obtained by bootstrapping, and for the second by likelihood theory. The estimators are nearly unbiased, and have for most practical cases smaller standard error than D. The likelihood-based estimator generally gives the highest precision. The relationship between fragment counts and precision is studied using simulation. The usual range of band counts (50–100) appears nearly optimal. The methodology is illustrated using data from a phylogenetic study on lettuce.

Communicated by M. Kearsey.

G. Gort (✉) · F. van Eeuwijk
Biometris, Wageningen University and Research Center, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands
e-mail: gerrit.gort@wur.nl

T. van Hintum
Centrum voor Genetisch Bronnen, Plant Research International, Wageningen University and Research Center, Bornsesteeg 65, 6708 PD Wageningen, The Netherlands

Introduction

AFLP is a DNA fingerprinting technique, that has been employed in many studies on plants (e.g. Tams et al. 2005), but also in studies on fungi (e.g. Mebrate et al. 2006), bacteria (e.g. Duim et al. 2001), and animals (e.g. Foulley et al. 2006). The resulting DNA fingerprints, also called profiles, are used in a wide spectrum of applications, like QTL studies (e.g. Zhong et al. 2006), diversity studies (e.g. van Berloo et al. 2008), and optimization of gene bank management (e.g. Jansen and van Hintum 2007). The question has been raised whether AFLP will remain useful in the near future, given the advances in genome sequencing, and new large-scale genotyping techniques like DArT (Wenzl et al. 2004). Meudt and Clarke (2007) suggest that fingerprinting techniques in general, and AFLP in particular, will remain valuable, especially if new analysis methods are developed, which overcome the problems arising in the analysis of AFLP data.

In this paper, we study the estimation of pairwise genetic similarity from dominant AFLP data. Estimation of similarity may be hampered by errors in, or erroneous interpretation of the binary band information from the AFLP profiles. As Bonin et al. (2007) mention, two types of errors prevail in AFLP genotyping: scoring errors and homoplasy. Many papers study the problem of scoring

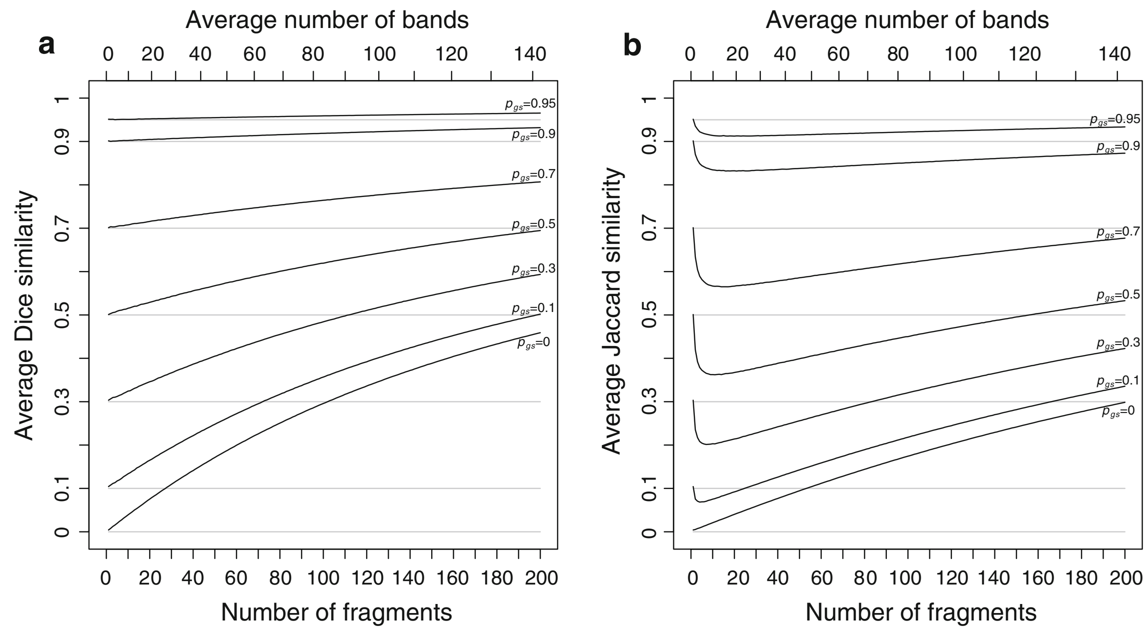

Fig. 1 **a** Average Dice, and **b** average Jaccard similarities as a function of number of fragments for 100,000 simulated pairs of profiles with genetic similarities $p_{gs} = 0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95$. Fragments are sampled from *fld* F_S with scoring range 51–500. The *top axes* show the average number of bands on a non-linear scale

errors (e.g. parts of Meudt and Clarke 2007, and papers cited therein), but here we focus on homoplasy.

Estimation of genetic similarity is biased due to size homoplasy, see Fig. 1 (to be discussed later in greater detail). Size homoplasy occurs if, for two individuals, equally sized, but different DNA fragments comigrate in two AFLP lanes, resulting in identical bands. The two bands are usually considered homologous. Hence, part of the observed similarity can be attributed to chance. Size homoplasy is considered to be one of the major problems in the analysis of AFLP data (Meudt and Clarke 2007; Robinson and Harris 1999). Caballero et al. (2008) study the effect of size homoplasy on estimates of genetic diversity and detection of selective loci. Empirical estimates of the amount of homoplasy can be found, e.g. in O'Hanlon and Peakall (2000), who report that among congeneric thistles comigrating fragments showed on average 2.5% size homoplasy, but among different subtribes up to 100%. Because of this problem, AFLP is commonly advised to be used only to assess relationships of closely related taxa (Althoff et al. 2007).

Another problem, related to the size homoplasy mentioned above, is the occurrence of two or more equally sized, but different fragments within a *single* lane. As two equally sized different fragments in two lanes generally comigrate, and are wrongly interpreted as homologous, they will also comigrate if amplified within a single lane, colliding in a single band, and wrongly interpreted as single fragment. We call the comigration of equally sized fragments within a single lane collision. In an empirical study on sugarbeet, Hansen et al. (1999) quantified the problem. They found 13.2% of the bands to contain collisions. In an in silico study of AFLP for a wide variety of species, Althoff et al. (2007) found fractions of bands containing collisions up to 49%, depending on the number of bands in a lane. Vekemans et al. (2002) reported in a Monte Carlo simulation study an average percentage of 30% of undetectable fragments. Collisions were studied from a probabilistic point of view in Gort et al. (2006) and Gort et al. (2008). Their theoretical results, which are at the basis of the present paper, are in line with the empirical results given above. Collisions also affect the estimation of genetic similarity.

Although it is recognized that both size homoplasy and collision may occur in AFLP, no attempts are usually made to correct for the problems: two equally sized bands are considered homologous, and a single band is interpreted as a single fragment. The reasons for this negligence are at least twofold: it is felt that the problems are minor (in the cases where AFLPs are suggested to be used), and hardly any methodology exists to correct for it. In Koopman and Gort (2004) a crude approach was proposed for the calculation of similarities from AFLP profiles.

In the present paper new estimators of genetic similarity from AFLP bands, corrected for homoplasy and collision, are proposed, one based on modification of the Dice and Jaccard coefficients, and one based on maximum likelihood. We take the following steps in the "Materials and methods" part to arrive at these estimators.

- We first review the AFLP procedure as a sampling method of DNA fragments.
- Next, the procedure and data are described from a modeling point of view, introducing notation, and a definition of pairwise genetic similarity for binary AFLP data is given.
- We review some commonly used similarity coefficients.
- We demonstrate, by simulation, that homoplasy and collision may seriously bias similarity estimates, resulting in Fig. 1.
- A first step towards a solution is to estimate the number of fragments in a lane from the number of bands. We describe two ways to do this, depending on the availability of band position information.
- Using estimated fragment counts, modified Dice (and Jaccard) coefficients in two versions are proposed, depending on availability of band position information.
- If band position information is available, a second estimator of genetic similarity is proposed, based on maximum likelihood (m.l.).
- Standard errors and confidence intervals are obtained, using the bootstrap for the modified coefficients, and standard likelihood theory for the m.l. estimator.
- Further distributional characteristics of the estimators are studied by simulation. We describe precisely how we sample AFLP profiles.

Using the m.l. estimator and its precision, we next focus on the question how many bands in a lane should be used to estimate genetic similarity optimally. The theory is illustrated by a small case study on lettuce, using data from a phylogenetic study by Koopman et al. (2001). Results of the simulations and the case study are shown in "Results". Conclusions are compiled and discussed in "Conclusions and discussion". The paper ends with appendices on bootstrapping and an overview of all symbols used.

Materials and methods

AFLP reviewed

To understand the ideas we are proposing, a short review of the AFLP fingerprinting technique is useful. The AFLP technique, developed by Keygene N. V. (Vos et al. 1995), can be looked upon as a sampling technique of DNA fragments from, hopefully, random locations within a genome. To arrive at a sample of DNA fragments representing an individual genome four steps are taken:

1. The total genomic DNA is cut into fragments by two restriction enzymes, often *Mse*I ("frequent cutter") and *Eco*RI ("rare cutter"). The result is a soup of fragments, flanked with restricted *Eco*RI–*Eco*RI, *Eco*RI–*Mse*I, or *Mse*I–*Mse*I sites.
2. Two adaptors, specific for the restriction enzymes, are ligated to the fragments, allowing primers to adhere in the third step.
3. Two primers, complementary to the two adaptors, with one or more selective nucleotides select a number of fragments for PCR amplification. In this way a *sample* of fragments is drawn. Primers with more selective nucleotides will select fewer fragments. If the four nucleotides A–C–T–G occur equally often in the genome, one extra selective nucleotide on, e.g. the *Eco*RI primer will cause a fourfold reduction in sample size of *Eco*RI–*Mse*I fragments, and a 16-fold reduction of the *Eco*RI–*Eco*RI fragments.
4. The amplified fragments are separated by length in a lane of a gel or capillary electrophoresis system. Shorter fragments travel further. Usually only fragments with at least one *Eco*RI primer are labeled, and will become visible as bands. Only fragments with lengths within a certain scoring range (e.g. 51–500 nucleotides long) are visualized as bands.

On a single gel multiple individual genomes are fingerprinted, one per lane. The lengths of the bands are determined by comparison with the position of DNA fragments of known lengths (sizers) in size ladders. For a complete review of the AFLP technique, see e.g. Mueller and LaReesa Wolfenbarger (1999).

AFLP modeled: single profile

In this section, we again step through the AFLP procedure, but now aim to statistically model the procedure and data. For convenience, we compile all introduced symbols in Appendix 2 (Table 7). We describe the procedure for a single lane of a gel.

In the first two steps of the procedure, the total genomic DNA is cut into fragments, and adaptors are ligated. Only part of these fragments are eligible for visualization: fragments containing at least one labeled site (e.g. *Eco*-RI site), and within the used scoring range (e.g. with 51–500 nucleotides) are candidates. We call this subset the *population* of fragments Π, containing, say, M fragments. Different restriction enzymes will result in different populations of fragments. The size and nucleotide composition of the genome also affect Π.

The length of a fragment is the number of nucleotides, adaptors included. We label the possible lengths of the fragments in Π with index i, ranging from 1 (referring to the smallest length in the scoring range) to N (referring to the largest length; e.g. with scoring range 51–500 $N = 450$). The probability distribution of the lengths is called the *fragment length distribution* (*fld*). With p_i the probability that a fragment, randomly drawn from Π, has length i, we can write $fld = (p_1, p_2,\ldots, p_N)$; note that $\sum_{i=1}^{N} p_i = 1$. Shorter fragments are more frequent than longer fragments, i.e. the *fld* is monotonically decreasing and skewed to the right (Gort et al. 2006). The amount of skewness is mainly determined by the GC content of the genome, if the frequent cutter *Mse*I is used. Lower GC content results in shorter fragments.

We assume the *fld* is known, or, at least, there is a reliable estimate of it. For all simulations we use *fld* F_S, estimated from the *Arabidopsis thaliana* genome based on in silico AFLP, as in Gort et al. (2006). This *fld* is reasonable for genomes with GC content close to 36%. For the estimation of the *fld* for other genomes we refer to the same publication.

In step 3 the primers select a *sample* of fragments from Π, selecting only those fragments, which have specific nucleotides next to the restriction sites. This resembles systematic sampling, but with unknown sample size. We treat the lengths of the sampled fragments as a random sample from *fld*. Assuming a constant but unknown sampling probability π for the fragments of Π, the number of fragments in the sample, called k, has approximately a Poisson distribution with expected count $m = \pi M$.

In step 4 the k fragments are separated by length, and visualized as bands. We assume that the position of a band within a lane is determined principally by the fragment length. Hence, a band will occur approximately at one of N discrete positions within a lane, which we call band lengths. A consequence is that two different fragments of the same length will occur as a single band.

The end product is a profile, containing bands at discrete positions, which can be represented by a binary vector $y = (y_1, y_2,\ldots, y_N)$. The binary variable y_i ($i = 1,\ldots, N$) indicates whether a band with length i is present. The number of bands in a lane is $n = \sum_{i=1}^{N} y_i$. Notice that the number of bands cannot be larger than the number of fragments ($n \le k$).

AFLP modeled: pairs of profiles and their similarity

Two related individuals share parts of their DNA. As a consequence, they share part of their two populations of fragments Π_1 and Π_2, containing M_1 and M_2 fragments, formed at step 2. This common part is called Π_a, and contains M_a fragments. The complement of Π_a within Π_1 is called Π_b, consisting of M_b fragments present in individual 1, but absent in 2. The complement of Π_a within Π_2 is called Π_c, and consists of M_c fragments, present in 2, but absent in 1. Π_b and Π_c are called the populations of unique fragments. Notice that $M_1 = M_a + M_b$, and $M_2 = M_a + M_c$. All population sizes M_a, M_b, and M_c are unknown. The fractions of common fragments are $F_1 = M_a/M_1$ and $F_2 = M_a/M_2$, which need not be the same, e.g. if the genomes have different sizes.

We define the pairwise genetic similarity p_{gs} of a pair of genotypes as the weighted average of fractions F_1 and F_2, with weights proportional to the population sizes:

$$p_{gs} = \frac{M_1}{M_1 + M_2} F_1 + \frac{M_2}{M_1 + M_2} F_2 = w_1 F_1 + w_2 F_2 \qquad (1)$$

Notice that p_{gs} can be written as $2\,M_a/(2\,M_a + M_b + M_c)$.

We assume that Π_a, Π_b, and Π_c have the same fragment length distribution *fld*.

In step 3 samples from *fld* are taken, resulting in sample sizes of fragments k_a, k_b, and k_c, approximately Poisson distributed with expected fragment counts m_a, m_b, and m_c, proportional to M_a, M_b, and M_c. The expected numbers of fragments of the two profiles are $m_1 = m_a + m_b$ and $m_2 = m_a + m_c$.

The end product after step 4 is a pair of profiles, represented by two binary band vectors $y_1 = (y_{11},\ldots, y_{N1})$, and $y_2 = (y_{12},\ldots, y_{N2})$, with band counts $n_j = \sum_{i=1}^{N} y_{ij}$ ($j = 1, 2$). We use the following notation for band counts:

a = number of shared bands in the two profiles $= \sum_{i=1}^{N} y_{i1} y_{i2}$;
b = number of bands in the first profile, but absent in the second $= \sum_{i=1}^{N} y_{i1}(1 - y_{i2})$;
c = number of bands in the second profile, but absent in the first $= \sum_{i=1}^{N} (1 - y_{i1}) y_{i2}$;
d = number of empty positions in both profiles $= \sum_{i=1}^{N} (1 - y_{i1})(1 - y_{i2})$.

Hence, a, b, c and d are the number of 1–1, 1–0, 0–1, and 0–0 matches, respectively. If more than two profiles are compared, d is often defined as the number of 0–0 matches in two lanes, limited to those band lengths with at least one band in one of the other lanes.

Commonly used similarity coefficients

We now review some commonly used similarity coefficients for binary AFLP data. From the similarity coefficients, reviewed by Reif et al. (2005), only the Dice, Jaccard's, and simple matching coefficient are relevant, because we treat AFLP as a dominant marker system.

The Dice coefficient (Dice 1945) D is an estimator of p_{gs}: $D = \frac{2a}{2a+b+c} = \hat{w}_1 \hat{F}_1 + \hat{w}_2 \hat{F}_2$ with weights $\hat{w}_1 = \frac{n_1}{n_1+n_2}$, $\hat{w}_2 = \frac{n_2}{n_1+n_2}$, and $\hat{F}_1 = \frac{a}{n_1}$, $\hat{F}_2 = \frac{a}{n_2}$. In genetic contexts the Dice similarity is often referred to as the Nei–Li similarity (Nei and Li 1979).

The Jaccard coefficient (Jaccard 1908) $J = \frac{a}{a+b+c}$ is the fraction of common bands compared to the total number of different bands for the two profiles. It is an estimator of

$M_a/(M_a + M_b + M_c)$, and not of the genetic similarity, as we define it. A non-linear relationship exists between J and D: $J = \frac{D}{2-D}$. For example, taking equal band counts in the two profiles: if half of the bands in each profile is shared, then $D = 1/2$, and $J = 1/3$. Examples of applications of Dice and Jaccard's coefficients as measures of genetic similarity are Drossou et al. (2004), and Tams et al. (2005).

The simple matching coefficient (Sneath and Sokal 1973) $S = \frac{a+d}{a+b+c+d}$ measures similarity including the 0–0 matches in the profiles as well, counting the 1–1 and 0–0 matches alike.

To illustrate the differences between the coefficients, take two genotypes with profiles containing 100 bands each, with $N = 450$, $a = 50$, $b = 50$, $c = 50$, hence $d = 300$. Since half of the bands of each profile is shared, $D = 0.5$, and $J = 0.33$, whereas $S = 0.78$. Suppose that for the same genotypes a second set of profiles is made, using primers with more selective nucleotides, and hence smaller samples of amplified fragments. Assuming a proportional decrease of band counts of 50% (so $a = 25$, $b = 25$, $c = 25$, and $d = 375$), we still have $D = 0.5$, and $J = 0.33$, but $S = 0.89$. Hence, S changes if the band counts decrease proportionally, whereas D and J remain constant.

Usually more than two genotypes are compared in a study. Often, for S only the 0–0 matches are counted for the occupied band positions in the whole set of genotypes. With a proportional decrease of the band counts a, b and c, the null count d will also decrease, but likely at a different rate. Hence, S will likely change, whereas D and J remain constant. S can also change if the set of other genotypes under study is changed. Wong et al. (2001) supply reasons in the realm of codominance of AFLP to avoid similarity measures exploiting 0–0 matches. Therefore, S has a number of undesirable properties. Only D is an estimator of pairwise genetic similarity, as we have defined it.

The problem: homoplasy and collision

To appreciate the possible consequences of homoplasy and collisions in relationship studies based on AFLP data, we performed a simulation study. We sampled 100,000 pairs of profiles for a range of genetic similarities p_{gs} (=0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95) and fragment counts $m_1 = m_2$ (=1,…, 200). The maximum fragment count $m = 200$ corresponds to $\approx$140 bands, which is about the maximum number of bands per lane to be found in practice. Each pair was sampled in three steps. First, a random draw k_a from the binomial (m_1, p_{gs}) distribution determined the sample size of fragments from the common part Π_a, the remaining $k_b = m_1 - k_a$ and $k_c = m_2 - k_a$ fragments to be sampled from the unique parts Π_b and Π_c. Next, k_a, k_b, and k_c lengths were sampled from the *fld*, and results were combined into two vectors of length $N = 450$, containing the counts of lengths 1,…, 450 for the two profiles. In the last step, a pair of binary vectors was created, containing absence/presence information of at least one fragment of length 1,…, 450, and representing a pair of AFLP profiles. Dice and Jaccard coefficients D and J were calculated for each pair, and averaged over all pairs to produce Fig. 1. The graph shows the average D and J as a function of the fragment count. The average band count is shown at the top axis on a non-linear scale. As an example, profiles with 100 fragments tend to produce approximately 83 bands, hence 17 collisions.

D overestimates the true genetic similarity seriously, increasingly so for larger fragment or band counts, and for smaller genetic similarities. For example, at band count 60 the average D has approximate biases 0.015, 0.085, and 0.23 for $p_{gs} = 0.9$, 0.5, and 0.0, respectively. At band count 100 the biases are 0.025, 0.14 and 0.34, respectively.

J is for band counts in the range 60,…, 100 sometimes lower than the true p_{gs} (if $p_{gs} > 0.3$), sometimes close to p_{gs} (if $p_{gs} \approx 0.3$) and sometimes higher (if $p_{gs} < 0.3$).

Estimation of number of fragments

The basic idea in this paper is that, in order to estimate genetic similarity, we need to know how many *fragments* from the two profiles are identical, whereas the profiles indicate how many *bands* are identical. The first step to solve this problem is to estimate the expected number of fragments m that gave rise to the n observed bands in a single profile. The difference between number of fragments and number of bands is called the collision count.

To estimate m, we discriminate between situations without and with band length information. Notice that band lengths are not always available, although in principle the information can be read from an AFLP gel, if size ladders are used. The lack of band length information is often based on limitations in the realm of intellectual property, as commercial players like Keygene N.V. propagate.

In the case of unknown band lengths, the collision count for a given *fld* is estimated from the band count, using Bayes' rule and generalized occupancy distributions, see Gort et al. (2006). The resulting estimator of the expected number of fragments m is called $\hat{m}_{\bar{L}}$.

With known band lengths, the number of collisions can be estimated using Bayes' rule and approximated multinomial tail probabilities, or applying the EM-algorithm, as in Gort et al. (2008). In the present paper, we report a simpler approach to arrive at an estimator of m. We propose a generalized linear model (g.l.m.) (McCullagh and Nelder 1991) for the binary band scores y_i. The scores y_i are assumed to be independent, and Bernoulli (P_i) distributed, with expected score $E(y_i) = P_i$ the probability that a band occurs with length i, if a sample of m fragments has been drawn from *fld*= $(p_1, \ldots, p_N)$. The band probability P_i and fragment probability p_i are related as: $(1 - P_i) = (1 - p_i)^m$, because the event "no band of length i" is equivalent with "none of the m fragments has length i". This equation can be transformed into

$$\log(-\log(1 - P_i)) = \log(m) + \log(-\log(1 - p_i)),$$

revealing the systematic part of the g.l.m. Hence, we fit a regression model for the band scores y_i, using $\log(m)$ as intercept, offset $\log(-\log(1 - p_i))$, and complementary log–log link. The estimator $\hat{m}_L$ of m is obtained by exponentiation of the estimator of the intercept $\log(m)$.

Modified Dice and Jaccard coefficients using binary AFLP data

Suppose we have two profiles with observed band counts $n_1 = a + b$, and $n_2 = a + c$. The expected numbers of fragments m_1 and m_2 are estimated by $\hat{m}_1$ and $\hat{m}_2$ by either of the two estimators from the previous section. The pairwise genetic similarity to be estimated is $p_{\mathrm{gs}} = \frac{M_1}{M_1+M_2}\frac{M_a}{M_1} + \frac{M_2}{M_1+M_2}\frac{M_a}{M_2} = w_1F_1 + w_2F_2$, as in (1).

For weights w_1 and w_2, we have straightforward estimators $\hat{w}_1 = \frac{\hat{m}_1}{\hat{m}_1+\hat{m}_2}$, and $\hat{w}_2 = \frac{\hat{m}_2}{\hat{m}_1+\hat{m}_2}$, since expected fragments counts are assumed to be proportional to population sizes. However, for the fractions common fragments $F_1 = \frac{M_a}{M_1}$ and $F_2 = \frac{M_a}{M_2}$, an estimator $\hat{m}_a$ of the number of common fragments m_a is needed. We estimate m_a as $\hat{m}_a = \hat{m}_1 + \hat{m}_2 - \hat{m}_{1+2}$, by analogy with the number of shared bands a, which can be calculated as $a = n_1 + n_2 - n_{1+2}$. In this formula $n_{1+2} = a + b + c$ is the total number of *different* bands, as if combining the two profiles into a single profile, and counting the bands. In the formula for $\hat{m}_a$, $\hat{m}_{1+2}$ is the estimated fragment count for the combination of the two profiles. The rationale of estimator $\hat{m}_a$ is the following: $\hat{m}_1$ estimates the number of fragments from the n_1 bands of profile 1, and $\hat{m}_2$ from the n_2 bands of profile 2. The sum $\hat{m}_1 + \hat{m}_2$ estimates the total number of fragments in the two lanes. Some of the fragments are counted twice, as they occur in both profiles. If we overlay profiles 1 and 2, we see what would have happened if we mixed the populations of fragments for the two genomes, and made a profile for the mixture. Identical fragments in the two populations, selected for amplification, will appear as a single band now, and $\hat{m}_{1+2}$ estimates the total number of fragments in the profile for the mixture, that is the number of *different* fragments in the mixture. Then the difference $(\hat{m}_1 + \hat{m}_2) - \hat{m}_{1+2}$ estimates $\hat{m}_a$, i.e. the number of fragments the two profiles have in common.

This results in $\hat{F}_1 = \frac{\hat{m}_a}{\hat{m}_1}$ and $\hat{F}_2 = \frac{\hat{m}_a}{\hat{m}_2}$. Estimators of unique fragment counts are $\hat{m}_b = \hat{m}_1 - \hat{m}_a$, and $\hat{m}_c = \hat{m}_2 - \hat{m}_a$.

As estimator of genetic similarity p_{gs} we now propose the modified Dice coefficient

$$D^{\mathrm{mod}} = \frac{\hat{m}_1}{\hat{m}_1 + \hat{m}_2}\frac{\hat{m}_a}{\hat{m}_1} + \frac{\hat{m}_2}{\hat{m}_1 + \hat{m}_2}\frac{\hat{m}_a}{\hat{m}_2} = \frac{2\hat{m}_a}{2\hat{m}_a + \hat{m}_b + \hat{m}_c},$$

replacing the band counts in the original Dice coefficient by estimated fragment counts. The Jaccard coefficient may be modified in the same way:

$$J^{\mathrm{mod}} = \hat{m}_a/(\hat{m}_a + \hat{m}_b + \hat{m}_c)$$

The maximum of both D^{mod} and J^{mod} is 1, occurring if the two profiles are identical. At the other end of the scale, there is no intrinsic limitation both for D^{mod} and J^{mod} to take on negative values, whereas $p_{\mathrm{gs}} \geq 0$. A solution to the problem is truncation of the estimator at 0.

The modified coefficients come in two versions, for situations without and with band length information. If band lengths are unknown, estimator $\hat{m}_{\bar{L}}$ is used, resulting in modified Dice and Jaccard coefficients

$$D^{\mathrm{mod}}_{\bar{L}} = 2\hat{m}_{\bar{L}a}/(2\hat{m}_{\bar{L}a} + \hat{m}_{\bar{L}b} + \hat{m}_{\bar{L}c}), \text{ and}$$
$$J^{\mathrm{mod}}_{\bar{L}} = \hat{m}_{\bar{L}a}/(\hat{m}_{\bar{L}a} + \hat{m}_{\bar{L}b} + \hat{m}_{\bar{L}c})$$

If band lengths are known, we use estimator $\hat{m}_L$, and the modified coefficients become

$$D^{\mathrm{mod}}_{L} = 2\hat{m}_{La}/(2\hat{m}_{La} + \hat{m}_{Lb} + \hat{m}_{Lc}), \text{ and}$$
$$J^{\mathrm{mod}}_{L} = \hat{m}_{La}/(\hat{m}_{La} + \hat{m}_{Lb} + \hat{m}_{Lc})$$

Maximum likelihood estimator of genetic similarity from binary AFLP data

In the case of known band lengths, a second estimator D^{mle} of the genetic similarity p_{gs} is proposed, based on maximum likelihood (m.l.) (Silvey 1975). For this estimator we need a statistical model for the data, consisting of the N pairs of binary scores $(y_{11}, y_{12}), (y_{21}, y_{22}), \ldots, (y_{N1}, y_{N2})$. We treat these pairs as independent. The two profiles have expected fragment counts $m_1 = m_a + m_b$ and $m_2 = m_a + m_c$, as before.

The four possible outcomes of a pair (y_{i1}, y_{i2}) are:

1. (0,0): no fragment of length i at all;
2. (0,1): no fragment from the unique part Π_b of genotype 1 and the common part Π_a, and at least one fragment from the unique part Π_c of genotype 2;
3. (1,0): at least one fragment from Π_b, and no fragment from Π_c and Π_a;
4. (1,1): either at least one fragment from Π_a, or at least one fragment from both Π_b and Π_c, but not from Π_a.

The likelihood of these four outcomes for the ith pair is:

1. (0,0): $\ell_i = (1 - p_i)^{m_b+m_a+m_c}$
2. (0,1): $\ell_i = (1 - p_i)^{m_b+m_a}(1 - (1 - p_i)^{m_c})$
3. (1,0): $\ell_i = (1 - (1 - p_i)^{m_b})(1 - p_i)^{m_a+m_c}$

4. (1,1): $\ell_i = \left(1-(1-p_i)^{m_a}\right) + \left(1-(1-p_i)^{m_b}\right)(1-p_i)^{m_a}\left(1-(1-p_i)^{m_c}\right)$

Next, the log-likelihood of the data $\mathrm{LL} = \sum_{i=1}^{N} \log(\ell_i)$ is maximized with respect to the parameters m_a, m_b, and m_c, resulting in m.l. estimators $\hat{m}_a$, $\hat{m}_b$, and $\hat{m}_c$. As in the previous section, we can define a modified Dice coefficient, now based on m.l. estimators, as

$$D_1^{\mathrm{mle}} = \frac{2\hat{m}_a}{2\hat{m}_a + \hat{m}_b + \hat{m}_c} = \hat{w}_1\hat{p}_1 + \hat{w}_2\hat{p}_2$$

with weights $\hat{w}_1 = \frac{\hat{m}_a+\hat{m}_b}{\hat{m}_a+\hat{m}_b+\hat{m}_a+\hat{m}_c}$, $\hat{w}_2 = \frac{\hat{m}_a+\hat{m}_c}{\hat{m}_a+\hat{m}_b+\hat{m}_a+\hat{m}_c}$, and $\hat{p}_1 = \frac{\hat{m}_a}{\hat{m}_a+\hat{m}_b}$, $\hat{p}_2 = \frac{\hat{m}_a}{\hat{m}_a+\hat{m}_c}$.

The m.l. procedure returns approximate standard errors of $\hat{m}_a$, $\hat{m}_b$, and $\hat{m}_c$, but not of D_1^{mle} as an estimator of p_{gs}. To get the precision of an estimator of p_{gs}, we reparameterize the likelihood. From $p_{\mathrm{gs}} = \frac{2M_a}{2M_a+M_b+M_c}$, it follows $\frac{p_{\mathrm{gs}}}{1-p_{\mathrm{gs}}} = \frac{M_a}{(M_b+M_c)/2} = \frac{m_a}{(m_b+m_c)/2}$, since we assume expected fragment counts proportional to population counts. Now, we replace m_a in the likelihood above by $\frac{p_{\mathrm{gs}}}{1-p_{\mathrm{gs}}}(m_b+m_c)/2$. Now the log-likelihood is maximized with respect to p_{gs}, m_b, and m_c, resulting in a direct m.l. estimator of p_{gs}, which we call D_2^{mle}.

A third parameterization replaces m_a by $\frac{1}{2}(m_b+m_c)\exp(l_{\mathrm{gs}})$, with $l_{\mathrm{gs}} = \mathrm{logit}(p_{\mathrm{gs}})$, yielding an estimator on the logit-scale, to be back-transformed to $D_3^{\mathrm{mle}}\,\mathrm{logit}^{-1}(\hat{l}_{\mathrm{gs}}) = \exp(\hat{l}_{\mathrm{gs}})/(1+\exp(\hat{l}_{\mathrm{gs}}))$. This estimator may have better distributional properties for p_{gs} close to 0 or 1.

Precision of the estimators

The precisions of estimators $D_{\bar{L}}^{\mathrm{mod}}$ and D_L^{mod} are determined by bootstrapping (Efron and Tibshirani 1993), whereas for D^{mle} the precision follows from standard likelihood theory.

For estimator $D_{\bar{L}}^{\mathrm{mod}}$ the following bootstrap method is used. The data for a pair of profiles consists of a pairs 1–1, b pairs 1–0, c pairs 0–1, and d pairs 0–0, collected in the vector (a, b, c, d), without knowledge of band lengths. For one bootstrap resample we take a sample of size N from the pairs 1–1, 1–0, 0–1, and 0–0, with probabilities given by a/N, b/N, c/N, and d/N, respectively. For this bootstrap sample the modified Dice coefficient is calculated as described, and stored.

For estimator D_L^{mod} a different bootstrap method is used. Now the band lengths are known. A bootstrap resample consists of a sample with replacement of N pairs (y_{i1}, y_{i2}) and connected *fld* probabilities p_i from the N pairs (y_{11}, y_{12}), $(y_{21}, y_{22}), \ldots, (y_{N1}, y_{N2})$, and a rescaling of the set of p_i's to have sum 1. Notice that the same pair (y_{i1}, y_{i2}), i.e. with the same band length, may occur more than once in the bootstrap resample. Therefore, a single bootstrap resample does not necessarily correspond to a pair of profiles, which could occur in practice. The method nevertheless works well, as shown later.

For $D_{\bar{L}}^{\mathrm{mod}}$ and D_L^{mod} we took 1,000 bootstrap samples, resulting in estimates of bias (defined as bootstrap mean − estimate), standard error, and bootstrap confidence intervals. We used accelerated bias-corrected percentile bootstrap confidence intervals, also known as BC_a confidence intervals (DiCiccio and Efron 1996). For a description of the calculation of these confidence intervals, as well as a comparison between different types of bootstrap confidence intervals, we refer to the appendix.

For estimator D_2^{mle} approximate standard errors follow from standard likelihood theory, leading to Wald confidence intervals for p_{gs} as $D_2^{\mathrm{mle}} \pm \mathrm{SE}(D_2^{\mathrm{mle}})z_{1-\alpha/2}$, with $z_{1\text{-}\alpha/2}$ the $1-\alpha/2$ quantile from the standard normal distribution. For D_3^{mle} we back-transform the Wald-confidence interval $\hat{l}_{\mathrm{gs}} \pm \mathrm{SE}(\hat{l}_{\mathrm{gs}})\cdot z_{1-\alpha/2}$ using logit^{-1}. Besides Wald-type confidence intervals we calculated profile likelihood confidence intervals for p_{gs} (see e.g. Venzon and Moolgavkar 1988). For profile likelihood confidence intervals the parameters m_b and m_c are treated as nuisance parameters, resulting in a profile likelihood for p_{gs} by maximizing over m_b and m_c.

Sampling of AFLPs and simulation

To study the behavior of the proposed estimators, we performed a simulation study. For a wide range of parameter settings (p_{gs}, m_1, and m_2) pairs of profiles were simulated by

1. calculating the expected counts of common fragments $m_a = ½(m_1 + m_2)p_{\mathrm{gs}}$, and unique fragments $m_b = m_1 - m_a$, and $m_c = m_2 - m_a$;
2. drawing random counts from Poisson distributions with means m_a, m_b, and m_c to arrive at fragment counts k_a, k_b, and k_c for the pair of profiles to be generated;
3. sampling separately k_a, k_b, and k_c fragment lengths from the *fld*;
4. combining the $k_a + k_b$ sampled fragments into the first profile, and $k_a + k_c$ fragments into the second, condensing the information into binary vectors y_1 and y_2 of length N.

For all combinations of p_{gs} = (0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95) and $m_1 = m_2$ = (40, 70, 120), we sampled 10,000 pairs of profiles. We also included a selection of unequal m's for some values of p_{gs}, to show that the methodology works in that case as well. For each pair of profiles the estimates $D_{\bar{L}}^{\mathrm{mod}}$, D_L^{mod} (with 1,000 bootstrap samples), and the three versions of D^{mle} were calculated.

Application of methodology: effect of number of fragments on precision

In AFLP profiling the number of fragments in a lane, and hence the number of bands, can be steered by the researcher by changing the number and/or type of selective nucleotides of the primers. Typical band counts per lane are between 50 and 100, corresponding to fragment counts from 60 to 125. The question arises whether these typical counts are optimal, i.e. whether the estimators of genetic similarity have highest possible precision.

In a simulation study we investigated for a number of examples (as before, $p_{gs} = 0.0$, 0.1, 0.3, 0.5, 0.7, 0.9, and 0.95 using $N = 450$ and *fld* F_S), how the standard error and width of the 95% profile likelihood confidence interval of p_{gs} based on D_2^{mle} depends on the fragment count. Expected fragment counts were varied from 15 to 500 (in steps of 5, equal expected counts for pairs of profiles), using 10,000 replicates at each step. We pushed the number of fragments to unrealistically high values now, to show the properties of D_2^{mle} in that case, at the same time realizing that in practice it is impossible to score profiles with very large numbers of bands per lane. In the simulations numbers of fragments up to 500 were allowed, resulting in profiles with more than 225 bands on average. In that case more than half of the band positions are occupied, since $N = 450$.

Case study: phylogenetic relations between Lactuca genera

The lettuce study by Koopman et al. (2001) aims at inferring species relationships in *Lactuca* and related genera from AFLP fingerprints. We selected one of the two primer combinations (E35/M49), and only 5 of the 20 species: *L. tenerrima*, *M. muralis, L. serriola*, *L. sativa,* and *L. tatarica*. We took 6–9 accessions for each of the five selected species. We selected the five species to have a wide range of band counts: mean counts (±SD) are 29.6 (±1.9), 32.4 (±2.5), 49.6 (±3.0), 52.6 (±2.8), and 84.1 (±5.1) for *L. tenerrima*, *M. muralis*, *L. serriola*, *L. sativa*, and *L. tatarica*, respectively.

For all pairs of accessions we calculated D, J, and D^{mle}. We used F_S from *A. thaliana* as *fld*. This seems reasonable, since the GC content of lettuce is close to that of *A. thaliana*: 36.6, 37, 38.2, 38.3, and 36.4% for the five species (Koopman et al. 2002) versus 36% for *A. thaliana*. The relationships between the species are visualized with UPGMA dendrograms, using dissimilarities $1 - D$, $1 - J$, and $1 - D^{mle}$.

Table 1 Average band counts n_1 and n_2, and Dice similarities D for 10,000 simulated pairs of AFLP profiles for a range of values of genetic similarity p_{gs} and expected numbers of fragments m_1 and m_2

Parameter settings			Results		
p_{gs}	m_1	m_2	n_1	n_2	D
0.0	40	40	37.0	37.0	0.1388
	70	70	61.3	61.2	0.2232
	120	120	96.4	96.3	0.3343
0.1	40	40	36.9	36.9	0.2192
	70	70	61.3	61.4	0.2936
	120	120	96.3	96.4	0.3902
0.3	40	40	36.9	37.0	0.3828
	70	70	61.3	61.3	0.4369
	120	120	96.4	96.4	0.5088
0.5	40	40	37.0	37.0	0.5522
	70	70	61.3	61.3	0.5870
	120	120	96.2	96.3	0.6355
0.7	40	40	36.9	36.9	0.7261
	70	70	61.2	61.1	0.7462
	120	120	96.4	96.3	0.7728
0.9	40	40	37.0	37.0	0.9061
	70	70	61.1	61.2	0.9131
	120	120	96.4	96.3	0.9213
0.95	40	40	37.0	37.0	0.9534
	70	70	61.3	61.3	0.9563
	120	120	96.4	96.4	0.9603
0.5	100	50	83.0	45.3	0.5736
	100	80	83.0	68.7	0.6057
0.7	70	40	61.3	37.0	0.7277
	80	70	68.7	61.3	0.7482

Fld F_S from *A. thaliana* is used, with $N = 450$ band positions

Results

General results from the simulation study

Table 1 shows some general results from the simulation study. For all simulation settings of p_{gs}, m_1, and m_2, the average band counts n_1, n_2, and average Dice similarity D are given. From the comparison of expected fragment counts with average band counts, we note that profiles with $m = 40$ have on average three collisions, with $m = 70$ on average 8.7 collisions, and with $m = 120$ on average 23.6 collisions. The ordinary Dice coefficient seriously overestimates the true similarity, with largest biases for small similarities and large fragment counts. The maximum observed bias is 0.334 for $p_{gs} = 0$ and $m = 120$. The smallest bias is 0.0034 for $p_{gs} = 0.95$ and $m = 40$.

Table 2 Results from a simulation study on D_L^{mod} for a range of values of genetic similarity p_{gs} and expected numbers of fragments m_1 and m_2, 10,000 replicated pairs of AFLP profiles, 1,000 bootstrap resamples, *fld* F_S from *A. thaliana* with $N = 450$

Parameter settings			Part I Results for D_L^{mod}				
p_{gs}	m_1	m_2	Mean and SE			95% BC_a bootstrap ci	
			Mean	Mean after bias correction	Bootstrap SE	Non-coverage % (too low, too high)	Length
0.0	40	40	−0.0016	−0.0014	0.0643	5.34 (3.04, 2.30)	0.2584
	70	70	−0.0028	−0.0022	0.0685	5.45 (2.88, 2.57)	0.2680
	120	120	−0.0030	−0.0024	0.0733	5.38 (3.11, 2.27)	0.2862
0.1	40	40	0.0986	0.0998	0.0743	4.53 (2.28, 2.25)	0.2942
	70	70	0.0987	0.0995	0.0712	4.93 (2.53, 2.40)	0.2781
	120	120	0.0970	0.0978	0.0717	4.92 (2.80, 2.12)	0.2797
0.3	40	40	0.2976	0.3002	0.0821	4.73 (2.16, 2.57)	0.3205
	70	70	0.2981	0.2997	0.0713	5.29 (2.55, 2.74)	0.2780
	120	120	0.2978	0.2989	0.0661	5.08 (2.58, 2.50)	0.2582
0.5	40	40	0.4976	0.5007	0.0788	4.30 (2.17, 2.13)	0.3070
	70	70	0.4974	0.4993	0.0653	4.72 (2.30, 2.42)	0.2548
	120	120	0.4977	0.4989	0.0576	4.99 (2.68, 2.31)	0.2250
0.7	40	40	0.6973	0.7000	0.0658	4.76 (2.47, 2.29)	0.2586
	70	70	0.6987	0.7003	0.0529	4.76 (2.41, 2.35)	0.2078
	120	120	0.6984	0.6993	0.0451	5.38 (2.73, 2.65)	0.1770
0.9	40	40	0.8978	0.8990	0.0391	3.83 (2.18, 1.65)	0.1613
	70	70	0.8994	0.9001	0.0309	4.29 (2.36, 1.93)	0.1250
	120	120	0.8996	0.9000	0.0258	4.65 (2.36, 2.29)	0.1032
0.95	40	40	0.9495	0.9501	0.0267	5.16 (1.59, 3.57)	0.1173
	70	70	0.9497	0.9500	0.0215	4.63 (2.22, 2.41)	0.0907
	120	120	0.9498	0.9500	0.0180	4.37 (2.30, 2.07)	0.0742
0.5	100	50	0.4975	0.4991	0.0599	5.03 (2.56, 2.47)	0.2336
0.5	100	80	0.4979	0.4993	0.0606	5.03 (2.65, 2.38)	0.2367
0.7	70	40	0.6979	0.6998	0.0551	4.81 (2.17, 2.64)	0.2157
0.7	80	70	0.6983	0.6998	0.0515	5.21 (2.83, 2.38)	0.2021
			Part II Results for truncated D_L^{mod}				
			Median and SE			95% BC_a bootstrap ci	
			Median	Median after bias correction	Bootstrap SE	Non-coverage % (too low, too high)	Length
0.0	40	40	0	0	0.0386	2.30 (2.30)	0.1522
	70	70	0	0	0.0374	2.57 (2.57)	0.1396
	120	120	0	0	0.0403	2.27 (2.27)	0.1387
0.1	40	40	0.0980	0.0992	0.0649	4.53 (2.28, 2.25)	0.2511
	70	70	0.0987	0.0998	0.0616	4.93 (2.53, 2.40)	0.2297
	120	120	0.0985	0.0997	0.0609	4.92 (2.80, 2.12)	0.2223
0.3	40	40	0.2985	0.3012	0.0818	4.73 (2.16, 2.57)	0.3197
	70	70	0.2990	0.3006	0.0712	5.29 (2.55, 2.74)	0.2776
	120	120	0.2974	0.2989	0.0660	5.08 (2.58, 2.50)	0.2579

Part I shows mean, mean after bias correction, mean of the bootstrap standard error, non-coverage percentage of 95% BC_a bootstrap confidence intervals (with left and right non-coverage percentages), and mean length of the interval. Part II shows, for $p_{gs} \leq 0.3$, the same type of results as part I, but for D_L^{mod} truncated at zero. Instead of means, medians are given. At $p_{gs} = 0.0$, only non-coverage at the right of $p_{gs} = 0.0$ is considered

Table 3 Results from a simulation study on D_L^{mod} for a range of values of genetic similarity p_{gs} and expected numbers of fragments m_1 and m_2, 10,000 replicated pairs of AFLP profiles, 1,000 bootstrap resamples, *fld* F_S from *A. thaliana* with $N = 450$

Parameter settings			Part I Results for D_L^{mod}				
			Mean and SE			95% BC_a bootstrap ci	
p_{gs}	m_1	m_2	Mean	Mean after bias correction	Bootstrap SE	Non-coverage % (too low, too high)	Length
0.0	40	40	−0.0009	−0.0008	0.0651	5.55 (3.15, 2.40)	0.2605
	70	70	−0.0017	−0.0014	0.0698	5.17 (2.68, 2.49)	0.2725
	120	120	−0.0021	−0.0015	0.0754	5.55 (2.99, 2.56)	0.2944
0.1	40	40	0.0989	0.1000	0.0749	4.52 (2.28, 2.24)	0.2957
	70	70	0.0996	0.1005	0.0721	5.05 (2.50, 2.55)	0.2815
	120	120	0.0978	0.0986	0.0733	5.17 (2.93, 2..24)	0.2861
0.3	40	40	0.2978	0.3004	0.0824	4.78 (2.34, 2.44)	0.3213
	70	70	0.2987	0.3003	0.0718	5.11 (2.43, 2.68)	0.2798
	120	120	0.2984	0.2995	0.0672	5.14 (2.56, 2.58)	0.2622
0.5	40	40	0.4977	0.5008	0.0789	4.38 (2.17, 2.21)	0.3075
	70	70	0.4978	0.4996	0.0655	4.80 (2.36, 2.44)	0.2558
	120	120	0.4982	0.4994	0.0582	5.26 (2.83, 2.43)	0.2275
0.7	40	40	0.6974	0.7001	0.0658	4.67 (2.43, 2.24)	0.2587
	70	70	0.6988	0.7003	0.0531	4.69 (2.41, 2.28)	0.2085
	120	120	0.6987	0.6997	0.0455	5.29 (2.51, 2.78)	0.1786
0.9	40	40	0.8979	0.8991	0.0391	3.78 (2.32, 1.46)	0.1618
	70	70	0.8994	0.9001	0.0309	4.28 (2.43, 1.85)	0.1253
	120	120	0.8997	0.9001	0.0259	4.62 (2.44, 2.18)	0.1040
0.95	40	40	0.9495	0.9501	0.0267	5.26 (1.74, 3.52)	0.1188
	70	70	0.9497	0.9500	0.0215	4.02 (2.21, 1.81)	0.0914
	120	120	0.9498	0.9500	0.0181	4.43 (2.34, 2.09)	0.0749
0.5	100	50	0.4978	0.4994	0.0600	5.19 (2.60, 2.59)	0.2342
0.5	100	80	0.4982	0.4997	0.0610	5.09 (2.69, 2.40)	0.2381
0.7	70	40	0.6981	0.6999	0.0551	4.97 (2.37, 2.60)	0.2160
0.7	80	70	0.6985	0.6999	0.0517	4.94 (2.71, 2.23)	0.2030
			Part II Results for truncated D_L^{mod}				
			Median and SE			95% BC_a bootstrap ci	
			Median	Median after bias correction	Bootstrap SE	Non-coverage % (too low, too high)	Length
0.0	40	40	0	0	0.0390	2.40 (2.40)	0.1524
	70	70	0	0	0.0395	2.49 (2.49)	0.1409
	120	120	0	0	0.0414	2.56 (2.56)	0.1419
0.1	40	40	0.0982	0.0992	0.0652	4.52 (2.28,2.24)	0.2510
	70	70	0.0997	0.1007	0.0621	5.05 (2.50,2.55)	0.2311
	120	120	0.1002	0.1011	0.0618	5.17 (2.93,2.24)	0.2249
0.3	40	40	0.2985	0.3013	0.0821	4.78 (2.34,2.44)	0.3204
	70	70	0.2999	0.3017	0.0716	5.11 (2.43,2.68)	0.2798
	120	120	0.2986	0.2997	0.0670	5.14 (2.56,2.58)	0.2618

Part I shows mean, mean after bias correction, mean of the bootstrap standard error, non-coverage percentage of 95% BC_a bootstrap confidence intervals (with left and right non-coverage percentages), and mean length of the interval. Part II shows, for $p_{\mathrm{gs}} \leq 0.3$, the same type of results as part I, but for D_L^{mod} truncated at zero. Instead of means, medians are given. At $p_{\mathrm{gs}} = 0.0$, only non-coverage at the right of $p_{\mathrm{gs}} = 0.0$ is considered

Results from the simulation study for modified Dice coefficients

Table 2 shows the results from the simulation study for the modified Dice coefficient $D_{\bar{L}}^{\mathrm{mod}}$, using profiles without band length information. In Table 3 results for D_{L}^{mod} are given. We notice the following.

1. Almost all of the bias of the original Dice coefficient is removed. $D_{\bar{L}}^{\mathrm{mod}}$ and D_{L}^{mod} slightly underestimate p_{gs} now (mean observed biases −0.0018 and −0.0015, averaged over all settings of p_{gs} and m, for $D_{\bar{L}}^{\mathrm{mod}}$ and D_{L}^{mod}, respectively), with largest observed bias equal to −0.0030 occurring for $D_{\bar{L}}^{\mathrm{mod}}$ in case $p_{\mathrm{gs}} = 0$ and $m = 120$. The remaining small negative bias can be removed even further by using a bootstrap bias correction. Mean observed biases are then −0.00058 and −0.00025.
2. The 95% (BC_a bootstrap) confidence intervals for the genetic similarity p_{gs} show reasonably good coverage properties. In 21 and 18 out of the 25 experimental settings the observed non-coverage is between 4.5 and 5.5%, hence deviations less than 0.5% from the nominal value of 5%. For both estimators the largest deviation from 5% is found for $p_{\mathrm{gs}} = 0.90$ and $m = 40$, with observed non-coverages of 3.8 and 3.8%, respectively. In these cases the confidence intervals are slightly too wide. For $p_{\mathrm{gs}} = 0.95$ and $m = 40$ the overall non-coverage behaves better (5.2 and 5.3%), but we find that in 1.6 and 1.7% of the cases the confidence intervals are too low, and in 3.6 and 3.5% too high, compared to the nominal 2.5 and 2.5%. In this case the intervals are too wide if the estimate is smaller than $p_{\mathrm{gs}} = 0.95$, and too narrow for estimates larger than 0.95.
3. The bootstrap standard errors of $D_{\bar{L}}^{\mathrm{mod}}$ and D_{L}^{mod} are smaller for larger number of expected fragments, with the exception of $p_{\mathrm{gs}} = 0$ and $p_{\mathrm{gs}} = 0.1$. Hence, in the examples for $p_{\mathrm{gs}} > 0.1$ larger fragment counts result in more precise estimates. The same can be said for the lengths of the 95% confidence intervals. If $p_{\mathrm{gs}} = 0.1$ the smallest standard error is observed for $m = 70$
4. The estimates $D_{\bar{L}}^{\mathrm{mod}}$ and D_{L}^{mod} may become negative for small values of p_{gs}. In the table this can be seen for $p_{\mathrm{gs}} = 0$, resulting in a negative average of D^{mod}, but it also occurs for $p_{\mathrm{gs}} = 0.1$. For $p_{\mathrm{gs}} = 0.3$ the lower bound of the 95% confidence interval may become negative. In practice a negative value of D^{mod} would be truncated at 0. Therefore, we added the bottom parts II of Tables 2 and 3, showing results for the truncated versions of $D_{\bar{L}}^{\mathrm{mod}}$ and D_{L}^{mod} for $p_{\mathrm{gs}} = 0.0$, 0.1, and 0.3. Since the truncation causes more distributional asymmetry we give medians instead of averages of $D_{\bar{L}}^{\mathrm{mod}}$ and D_{L}^{mod}. For $D_{\bar{L}}^{\mathrm{mod}}$ the bias-correction decreases the bias, but this is not always the case for D_{L}^{mod}. For $p_{\mathrm{gs}} = 0$ we give the non-coverage of the (97.5%) confidence interval only at the right of $p_{\mathrm{g}} = 0$. For $p_{\mathrm{gs}} = 0$ we observe the largest standard errors for the cases with largest m, suggesting that the optimal number of fragments is smaller than $m = 120$.
5. In all cases $D_{\bar{L}}^{\mathrm{mod}}$ has narrower 95% confidence intervals than D_{L}^{mod}, although differences are small (average difference in length is only 0.0019). In all cases the bootstrap $\mathrm{SE}(D_{\bar{L}}^{\mathrm{mod}}) \leq \mathrm{SE}(D_{L}^{\mathrm{mod}})$, but again differences are small. The coverage of the 95% confidence interval of $D_{\bar{L}}^{\mathrm{mod}}$ is slightly better than that of D_{L}^{mod}: average absolute deviation from the nominal 5% is 0.33% for $D_{\bar{L}}^{\mathrm{mod}}$ compared to 0.36% for D_{L}^{mod}. Intuitively better behavior of D_{L}^{mod} was expected, since D_{L}^{mod} exploits band length information, but we conclude, surprisingly, that $D_{\bar{L}}^{\mathrm{mod}}$ has slightly better characteristics than D_{L}^{mod}

Results from the simulation study for maximum likelihood estimators D^{mle}

Table 4 shows the results from the simulation study for D^{mle}. We notice the following.

1. Estimators D_1^{mle}, D_2^{mle}, and D_3^{mle} almost always return the same estimate. Only for $p_{\mathrm{gs}} \geq 0.9$ we see minor differences, resulting in means differing in the fourth decimal. Hence, only results for D_2^{mle} are shown.
2. The large positive bias of the original Dice coefficient is removed. For $p_{\mathrm{gs}} > 0.1$, a negligible negative bias of D_2^{mle} remains: the mean bias is −0.0015. For $p_{\mathrm{gs}} \leq 0.1$ a small positive bias is observed, because of the necessarily non-negative value of the estimators. For $p_{\mathrm{gs}} = 0$ the medians (not shown) are 0, and for $p_{\mathrm{gs}} = 0.1$ they are 0.0965 ($m = 40$), 0.0982 ($m = 70$), and 0.0995 ($m = 120$).
3. The 95% Wald confidence intervals for p_{gs} are conservative for small values of p_{gs} (non-coverage rates smaller than nominal value), but are becoming more and more liberal for larger values. Obviously, the approximate standard error of D_2^{mle} is too large for small values of D_2^{mle}, and too small for large values. The deviations from 5% seem acceptable for $0.3 \leq p_{\mathrm{gs}} \leq 0.7$ and $m > 40$. The number of intervals with a lower bound larger than the true p_{gs} outnumber those with an upper bound smaller than p_{gs}. This is also an indication of standard errors which are too high for low values of the estimate, and too small for large values.
4. The 95% profile likelihood confidence intervals for p_{gs} have for a large number of settings non-coverage rates

Table 4 Results from a simulation study on D^{mle} for a range of values of genetic similarity p_{gs} and expected numbers of fragments m_1 and m_2, 10,000 replicated pairs of AFLP profiles, 1,000 bootstrap resamples, *fld* F_S from *A. thaliana* with $N = 450$

Parameter settings			Results for D^{mle}							
			D_2^{mle}		Wald ci		Profile likelihood ci		Back transformed Wald ci	
p_{gs}	m_1	m_2	Mean	SE	Non-coverage% (too low, too high)	Length	Non-coverage% (too low, too high)	Length	Non-coverage % (too low, too high)	Length
0.0	40	40	0.0202	0.0759	0.59 (0.59)	0.1689	1.98 (1.98)	0.1401	–	–
	70	70	0.0203	0.0651	1.18 (1.18)	0.1477	2.35 (2.35)	0.1275	–	–
	120	120	0.0216	0.0611	1.48 (1.48)	0.1409	2.26 (2.26)	0.1236	–	–
0.1	40	40	0.1004	0.0721	2.41 (0.97, 1.44)	0.2320	4.49 (2.40, 2.09)	0.2364	5.28 (0, 5.28)	0.4532
	70	70	0.1006	0.0645	3.13 (1.25, 1.88)	0.2144	4.64 (2.29, 2.35)	0.2164	5.51 (0, 5.51)	0.3974
	120	120	0.1001	0.0611	3.06 (0.94, 2.12)	0.2056	4.94 (2.59, 2.35)	0.2059	5.80 (0, 5.80)	0.3742
0.3	40	40	0.2979	0.0807	6.27 (3.56, 2.71)	0.3151	5.23 (2.70, 2.53)	0.3074	3.56 (0.02, 3.54)	0.3123
	70	70	0.2987	0.0690	5.77 (2.81, 2.96)	0.2702	5.24 (2.52, 2.72)	0.2660	3.81 (0.15, 3.66)	0.2670
	120	120	0.2985	0.0622	5.43 (2.69, 2.74)	0.2436	5.08 (2.63, 2.45)	0.2411	3.90 (0.24, 3.66)	0.2410
0.5	40	40	0.4978	0.0777	5.54 (2.09, 3.45)	0.3045	4.73 (2.42, 2.31)	0.2948	4.09 (1.40, 2.69)	0.2955
	70	70	0.4978	0.0643	5.29 (2.06, 3.23)	0.2519	4.90 (2.36, 2.54)	0.2495	4.38 (1.51, 2.87)	0.2467
	120	120	0.4981	0.0560	5.08 (2.21, 2.87)	0.2195	4.99 (2.74, 2.25)	0.2183	4.36 (1.76, 2.60)	0.2161
0.7	40	40	0.6974	0.0646	6.24 (1.66, 4.58)	0.2532	6.39 (3.56, 2.83)	0.2363	4.72 (2.35, 2.37)	0.2492
	70	70	0.6989	0.0523	5.53 (1.75, 3.78)	0.2049	5.01 (2.55, 2.46)	0.2030	4.56 (2.35, 2.21)	0.2028
	120	120	0.6987	0.0445	5.67 (1.65, 4.02)	0.1744	5.20 (2.43, 2.77)	0.1741	4.88 (2.20, 2.68)	0.1731
0.9*	40	40	0.8976	0.0382	7.74 (0.70, 7.04)	0.1496	12.60 (9.95, 2.65)	0.1301	3.77 (3.04,0.73)	0.1582
	70	70	0.8995	0.0304	6.96 (0.87, 6.09)	0.1192	7.76 (5.14, 2.62)	0.1092	4.24 (2.88, 1.36)	0.1238
	120	120	0.8997	0.0255	6.83 (1.05, 5.78)	0.0999	5.30 (2.50, 2.80)	0.0990	4.54 (2.87, 1.67)	0.1026
0.95*	40	40	0.9491	0.0266	13.42 (0.22, 13.20)	0.1007	19.34 (15.61, 3.73)	0.0899	6.43 (2.95, 3.48)	0.1199
*	70	70	0.9496	0.0208	9.96 (0.46, 9.50)	0.0817	12.87 (9.48, 3.39)	0.0736	3.73 (3.20, 0.53)	0.0923
*	120	120	0.9500	0.0175	9.06 (0.75, 8.31)	0.0684	6.47 (3.42, 3.05)	0.0659	4.10 (3.08, 1.02)	0.0750
0.5	100	50	0.4977	0.0592	5.56 (2.35, 3.21)	0.2320	5.23 (2.66, 2.57)	0.2301	4.71 (1.86, 2.85)	0.2280
0.5	100	80	0.4982	0.0595	5.36 (2.30, 3.06)	0.2330	4.88 (2.57, 2.31)	0.2314	4.54 (1.84, 2.70)	0.2289
0.7	70	40	0.6980	0.0544	5.96 (1.67, 4.29)	0.2132	5.65 (2.77, 2.88)	0.2054	4.97 (2.34, 2.62)	0.2109
0.7	80	70	0.6985	0.0509	5.75 (2.62, 2.21)	0.1995	5.20 (2.84, 2.36)	0.1983	4.83 (2.62, 2.21)	0.1976

Shown are the mean, mean standard error, and properties of three types of confidence intervals: non-coverage percentage (with left and right non-coverage percentages), and mean length of (1) 95% Wald c.i., (2) 95% profile likelihood c.i., and (3) 95% logit-back transformed Wald c.i. At $p_{gs} = 0.0$, only non-coverage at the right of $p_{gs} = 0.0$ is considered

* In case $p_{gs} = 0.9$ ($m = 40$), or $p_{gs} = 0.959$ ($m = 40, 70, 120$) identical pairs of profiles were sampled (10, 348, 53, and 10 times, respectively); in these cases $D_2^{mle} = 1$, with standard error 0, and we took logit(p_c) = 16 with standard error 0

close to 5%. In 16 out of the 25 settings the deviation of the non-coverage rate from the nominal value is less than 0.5%. Larger deviations are found for larger values of p_{gs} and smaller fragment counts. The largest deviation is observed for $p_{gs} = 0.95$ and $m = 40$, with a non-coverage rate equal to 19%, making the profile likelihood interval useless in this situation. The number of intervals with an upper bound smaller than p_{gs} becomes exceedingly large in these cases. The profile likelihood intervals work well for $p_{gs} < 0.7$, irrespective of the studied fragment counts, and for larger values of p_{gs}, but only if the fragment count is large enough.

5. The 95% back-transformed (from logit-scale) Wald confidence intervals generally have a non-coverage rate close to the nominal 5%. However, for small values of p_{gs} they are highly asymmetrically distributed (with respect to p_{gs}). Intervals with lower bounds exceeding p_{gs} dominate in these cases. If $p_{gs} = 0$, estimates of p_{gs} on the logit scale tend to $-\infty$, and the approximate standard errors are badly determined, resulting in useless confidence intervals. For high values of p_{gs}, intervals with upper bounds lower than p_{gs} get the upper hand. The back-transformed Wald confidence intervals are usable for $p_{gs} \geq 0.5$, and tend to be conservative then.

6. The standard error of D_2^{mle} decreases with larger expected fragment counts, as expected. For all three types of confidence intervals larger numbers of fragments result in narrower confidence intervals.
7. None of the three types of confidence intervals are usable for all values of p_{gs}. The profile likelihood intervals have the broadest range of application of p_{gs}:$p_{\text{gs}} < 0.7$ irrespective of m, and $p_{\text{gs}} \geq 0.7$ for larger values of m. The back-transformed Wald intervals perform best for large values of p_{gs}. The Wald confidence intervals are widest (at $p_{\text{gs}} = 0.5$ and 0.7), making them the least attractive in this range.
8. For all cases with $p_{\text{gs}} \geq 0.3$, D_2^{mle} has smaller standard errors than $D_{\bar{L}}^{\text{mod}}$ and D_L^{mod}. Furthermore, in all cases the profile likelihood confidence intervals based on D_2^{mle} are narrower than the bootstrap confidence intervals based on $D_{\bar{L}}^{\text{mod}}$ and D_L^{mod}. These results suggest that D_2^{mle} is to be preferred over the modified coefficients $D_{\bar{L}}^{\text{mod}}$ and D_L^{mod}.

Comparing standard errors

The simulation study has shown that the proposed estimators are approximately unbiased. Although attractive in itself, unbiasedness does not guarantee a higher precision since $\text{SE} = \sqrt{\text{bias}^2 + \text{var}}$. Using the data from the simulation study, we estimated bias(D), and var(D) by bootstrapping, and compared SE(D) with $\text{SE}\left(D_L^{\text{mod}}\right)$.

For most cases we find $\text{SE}(D) > SE\left(D_L^{\text{mod}}\right)$, with the most extreme outcome for $p_{\text{gs}} = 0.0$ and $m = 120$, where SE(D) is $4.5 \times \text{SE}\left(D_L^{\text{mod}}\right)$. For large values of p_{gs} ($p_{\text{gs}} = 0.95$, all m; $p_{\text{gs}} = 0.9$, $m = 40{,}70$; $p_{\text{gs}} = 0.7$, $m = 40$), we find that $\text{SE}(D) < \text{SE}\left(D_L^{\text{mod}}\right)$, but SE($D$) is never smaller than $0.95 \times \text{SE}\left(D_L^{\text{mod}}\right)$. Hence, depending on the combination of p_{gs} and m, very large gains in standard error can be obtained, or, for large p_{gs} (in combination with small fragment counts) minor losses. In the last cases, the gain in bias is outweighed by the loss in variance, and the new estimator D_L^{mod} is marginally less precise compared to D.

Results for the effect of expected number of fragments on precision

Figure 2 shows the results of the simulation study on the relationship between the expected number of fragments m and precision of D_2^{mle}. In the left-hand side figure the expected number of fragments is plotted against the average standard error of D_2^{mle}. At the top axis the average band count is shown. We observe the following:

1. Starting at small numbers of fragments, the standard error of D_2^{mle} decreases as the number of fragments increases. The rate of change of the standard error is

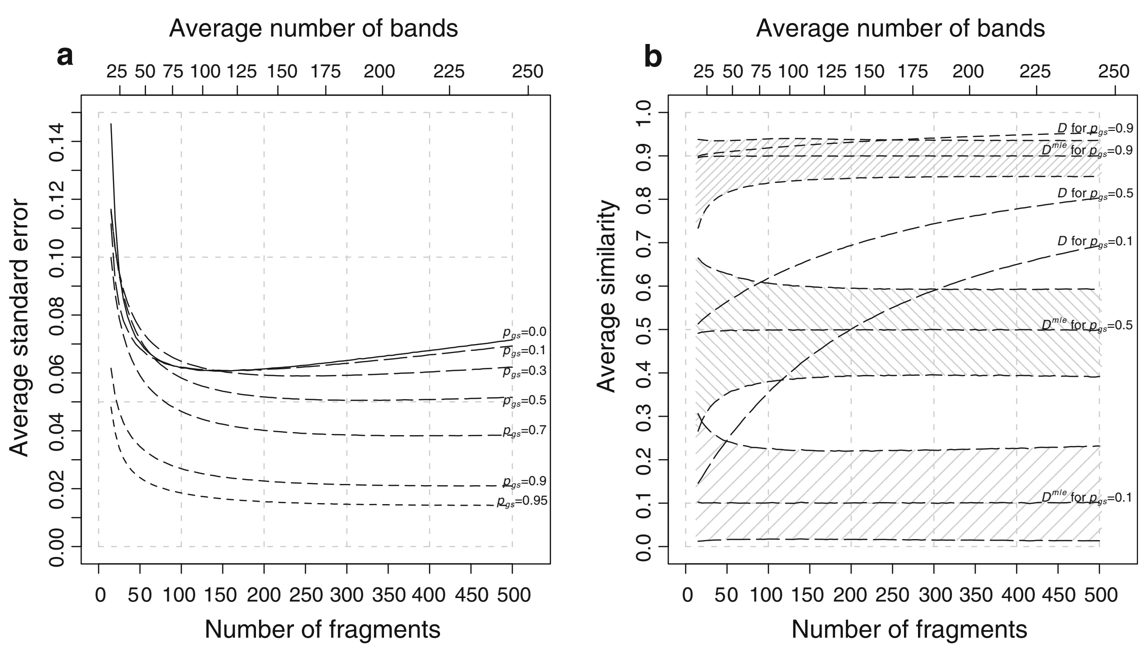

Fig. 2 **a** Average SE of D_2^{mle}, and **b** average D_2^{mle} and D, as functions of numbers of fragments for different values of p_{gs}. In plot **a** interpolated lines are drawn for fragment counts ranging from 15 to 500 in steps of 5 for $p_{\text{gs}} = 0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95$, and in plot **b** for $p_{\text{gs}} = 0.1$, 0.5, and 0.9. The *shaded areas* in **b** indicate average 95% profile likelihood confidence intervals of p_{gs}. For each value of m and p_{gs}, 10,000 pairs of profiles were sampled from *fld* F_S with scoring range 51–500. The *top axes* show the average number of bands on a non-linear scale

high at low fragment counts, but decreases. As the number of fragments increases, the standard error reaches a minimum, and afterwards increases again.

2. The optimal number of fragments depends on p_{gs}. Smaller values of p_{gs} allow smaller numbers of fragments. For $p_{gs} = 0$ or 0.1 the optimal number of fragments is close to $m = 140$ (or $n = 110$ bands). For $p_{gs} = 0.3$ this count is approximately $m = 250$ ($n = 165$), for $p_{gs} = 0.5$ $m = 350$ ($n = 205$), and for $p_{gs} = 0.7$ $m = 500$ ($n = 245$). For $p_{gs} = 0.9$ or 0.95 the optimal fragment count is larger than 500 fragments.
3. In general a large range of near-optimal fragment counts exists.
4. The usual range of band counts (between 50 and 100) is not optimal, especially if the focus is on highly related species with high p_{gs}. However, the gain in accuracy will generally be small if larger band counts are used. The small gain in accuracy must be balanced against the possible scoring problems that may occur with large band counts.

In the right-hand side figure the expected number of fragments is plotted against the average D_2^{mle}, and average Dice similarity. Furthermore, the average lower and upper bounds of the 95% profile likelihood confidence intervals are shown. For clarity, only results for $p_{gs} = 0.1$, 0.5, and 0.9 are given. We observe the following:

5. D_2^{mle} is an (almost) unbiased estimator of p_{gs}, even for extremely large fragment counts. For very small fragment counts ($m \leq 25$) there appears to be small negative bias.
6. Starting at small m, the width of the confidence interval quickly decreases. For large enough m (depending on p_{gs}) the width remains approximately constant.
7. The usual range of band counts, although not optimal, seems reasonable. Only little gain in the width of the confidence intervals can be expected from higher fragment counts, as in 4).
8. The confidence intervals are rather wide. The only way to reach narrower intervals is to use multiple gels with different primer combinations, and combine the information from the different profiles.

Results for case study on lettuce and related genera

Figure 3 shows the UPGMA dendrograms for the five species, split out for the three dissimilarity measures. The dendrograms for $1 - D$ and $1 - J$ are largely the same. With all three dissimilarities the species are separated well. Notice that the $1 - D^{\mathrm{mle}}$ dissimilarities are closer to 0, as expected. Notice further that the $1 - D^{\mathrm{mle}}$ dissimilarities are not a simple shift towards 1. In the hierarchical clustering scheme for D and J, *L. tenerrima* joins after clustering of *L. serriola*, *L. sativa*, and *L. tatarica*, but for D^{mle} *L. tenerrima* joins after clustering of *L. serriola* and *L. sativa* only. Apparently, *L. tenerrima* and *L. tatarica* have switched places. This behavior can be understood from the band count. The AFLP profiles for *L. tenerrima* contain a small number of bands, whereas *L. tatarica* profiles have large counts. Hence, bias corrections for comparisons with *L. tenerrima* are smaller than those with *L. tatarica*.

Conclusions and discussion

In this study, we propose new estimators of pairwise genetic similarity p_{gs} from binary AFLP data, correcting for homoplasy. We define pairwise genetic similarity for

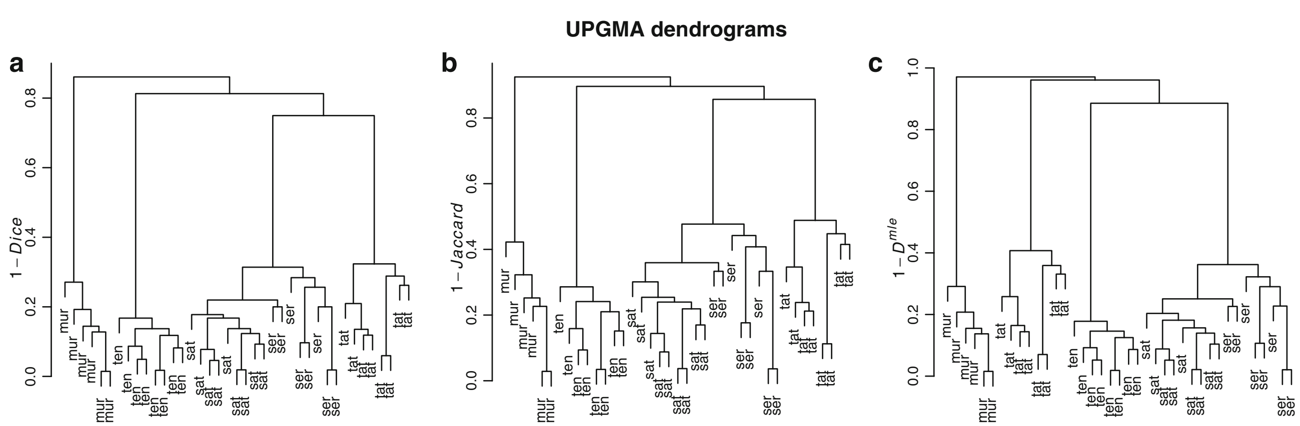

Fig. 3 UPGMA dendrograms for three dissimilarities: **a** 1−Dice D, **b** 1−Jaccard J, and **c** $1-D^{\mathrm{mle}}$ for five species of *Lactuca* and related genera, with 6–9 accessions per species. Labels are: ten = *L. tenerrima*, mur = *M. muralis*, ser = *L. serriola*, sat = *L. sativa*, and tat = *L. serriola*. For D^{mle} we used *fld* F_S with scoring range 110–501

AFLP data as the weighted average of fractions of common fragments. Using this definition, the Dice coefficient is a natural candidate for replacement, but a homoplasy corrected version of the Jaccard coefficient is suggested as well. For most practical cases the new estimators are better than the ordinary Dice coefficient, because the bias is removed, at the cost of a small increase in variance. Only for large genetic similarities in combination with low band counts (roughly: $p_{gs} = 0.95$ and $n < 100$, $p_{gs} = 0.90$ and $n < 65$, $p_{gs} = 0.70$ and $n < 38$), Dice performs better.

For profiles without band length information, we propose the modified Dice coefficient $D_{\bar{L}}^{mod}$. Using the bootstrap, standard errors and confidence intervals are obtained. The bootstrap allows a further reduction of the already small negative bias of $D_{\bar{L}}^{mod}$.

For AFLP profiles with band length information, we have three candidate estimators: D^{mle}, D_L^{mod}, and $D_{\bar{L}}^{mod}$. Best results are obtained using the maximum likelihood estimator D^{mle}, although differences are small. Second best is, surprisingly, $D_{\bar{L}}^{mod}$, ignoring the band length information. The standard error of D^{mle} follows from likelihood theory, hence no bootstrapping is needed. Profile likelihood confidence intervals for p_{gs} are narrowest. However, care has to be taken in the choice of type of confidence interval. Profile likelihood intervals are only acceptable, if $p_{gs} < 0.7$ irrespective of the number of fragments, and for $p_{gs} \geq 0.7$ if the fragment counts are large enough. For small fragment counts and large p_{gs}, more acceptable results are obtained for the back transformed Wald intervals, using an estimator on the logit-scale.

The modified Dice coefficients $D_{\bar{L}}^{mod}$ or D_L^{mod} are good alternatives as well. Over the whole range of p_{gs} the confidence intervals based on D_L^{mod} and $D_{\bar{L}}^{mod}$ showed more stable coverage properties than those for D^{mle}.

The homoplasy corrected estimate of genetic similarity is always smaller than the ordinary Dice coefficient, because part of the observed band similarity is attributed to chance. The magnitude of this correction depends on the true genetic similarity, but also on the fragment counts. Both smaller similarities and larger numbers of fragments lead to larger corrections.

The standard error of the similarity estimator D^{mle} and the width of the confidence interval cannot be made arbitrarily small by increasing the number of fragments in the profiles. The optimal number of fragments exists, but its value depends on the true genetic similarity, and there is a large range of near-optimal fragment counts. The usual range of band counts (between 50 and 100) is suboptimal, but in general the gain in precision is small if higher numbers of fragments are used, and should be balanced against increasing scoring problems.

To get more precise estimates of genetic similarity, multiple gels with different primer combinations or restriction enzymes should be used, and the information from the different profiles should be combined. D^{mle} can easily be modified to estimate a single genetic similarity from multiple pairs of profiles, even allowing for possibly different *fld*'s for the different profiles. Modifications of this type (beyond ordinary averaging) are less straight forward for the modified coefficients D_L^{mod} and $D_{\bar{L}}^{mod}$. This flexibility is a further argument in favor of D^{mle}.

To account for homoplasy and collisions properly, *all* bands in the profiles must be scored, not just the non-monomorphic bands. The effect of scoring non-monomorphic bands only is that Dice and Jaccard coefficients are lowered in a way that depends on the set of individuals under study. Inclusion or exclusion of a less related individual in the study, could result in exclusion or inclusion of bands, which are polymorphic with the individual, but monomorphic without. Hence, the similarity coefficient would be different with or without this individual.

Conclusions drawn here are mainly based on a single simulation study. Furthermore, we have to rely on a number of assumptions. For instance, we assume to know the *fld*, which in reality hardly ever is the case. Only if full DNA sequence information is available and by using in-silico AFLP procedures, do we have an estimate of the *fld* very close to the true *fld*. In other cases, a less reliable estimate of the *fld* may come from the GC content or directly from the binary AFLP data, as described in Gort et al. (2006).

Another topic related to the *fld*, is the fact that two distantly related individuals, e.g. with highly different GC contents, may have different *fld*s. In this paper we have assumed that there is a common *fld*. Further study on the effect of misspecification of the *fld*s on the statistical properties of the proposed estimators is needed.

In the present paper, we studied the effect of homoplasy and collision on the estimation of genetic similarity from binary AFLP data. Examples of studies that may directly benefit from the proposed homoplasy corrected estimates of genetic similarity are studies on genetic diversity, e.g. in plant genetic resources or breeding programs, but also phylogenetic and taxonomic studies, and studies of essential derivation, in which plant breeders try to establish thresholds for genetic similarity between initial and new, allegedly derived varieties (Van Eeuwijk and Law 2004).

In other studies where AFLP profiles are analyzed, the problem of homoplasy may have an impact as well. For example, in linkage studies for tracing quantitative trait loci (QTLs) or for mapping purposes, a band is interpreted as a single DNA fragment, residing at one unique locus of the genome. Here the best strategy may be to avoid homoplasy as much as possible, by limiting the number of fragments per lane, or avoiding bands corresponding to short fragments.

In population genetic applications of AFLP, homoplasy and collision may also affect estimation of parameters. For example, if the allele frequency of the DNA fragment corresponding to a band is the parameter of interest, like in Krauss (2000), who tested three procedures for estimation of null allele frequencies, homoplasy may cause some bands to be non-homologous, thereby changing the relative frequency of absent bands. Derived quantities like heterozygosity, coefficient of co-ancestry, or genetic distances, may need corrections for homoplasy and/or collision as well. These corrections require careful consideration, and are beyond the scope of the present paper. An example of a recent study of homoplasy in population genetics is Caballero et al. (2008), who focus on population genetic diversity and detection of selective loci.

In a study by Holland et al. (2008) about automated scoring of AFLPs, the suggestion is made to decrease the bin width for scoring fragments on a capillary system. This is another route towards a solution of the homoplasy problem, because the resulting profiles will likely have less homoplasy, albeit at the cost of an increased error rate for homologous fragments. In future work this approach may be joined with ours to arrive at improved evaluation of homoplasy.

The problem of homoplasy described here is not limited to the AFLP marker system. In a study on homology among RAPD fragments for three very closely related

Table 5 Comparison of bootstrap confidence intervals for p_{gs} from a simulation study on D_L^{mod} for a range of values of genetic similarity p_{gs} and expected numbers of fragments m_1 and m_2, 10,000 replicated pairs of AFLP profiles, 1,000 bootstrap resamples, *fld* F_S from *A. thaliana* with $N = 450$

Parameter settings			Results for D_L^{mod}: 95% bootstrap confidence intervals for p_{gs}					
			Percentile bootstrap c.i.		Bias corrected bootstrap c.i.		BC$_a$ c.i.	
p_{gs}	m_1,	m_2	Non-coverage% (too low, too high)	Length (trunc)	Non-coverage% (too low, too high)	Length (trunc)	Non-coverage% (too low, too high)	Length (trunc)
0.0	40	40	6.98 (5.57, 1.41)	0.2495 (0.1324)	6.50 (4.73, 1.77)	0.2517 (0.1386)	5.34 (3.04, 2.30)	0.2584 (0.1522)
	70	70	5.68 (3.63, 2.05)	0.2670 (0.1332)	5.60 (3.40, 2.20)	0.2672 (0.1354)	5.45 (2.88, 2.57)	0.2680 (0.1396)
	120	120	5.40 (3.08, 2.32)	0.2862 (0.1382)	5.48 (3.23, 2.25)	0.2862 (0.1381)	5.38 (3.11, 2.27)	0.2862 (0.1387)
0.1	40	40	5.70 (4.16, 1.54)	0.2893 (0.2369)	5.21 (3.47, 1.74)	0.2904 (0.2416)	4.53 (2.28, 2.25)	0.2942 (0.2511)
	70	70	5.25 (3.18, 2.07)	0.2777 (0.2255)	5.28 (3.01, 2.27)	0.2778 (0.2778)	4.93 (2.53, 2.40)	0.2781 (0.2297)
	120	120	5.05 (2.95, 2.10)	0.2796 (0.2220)	4.98 (2.86, 2.12)	0.2797 (0.2220)	4.92 (2.80, 2.12)	0.2797 (0.2223)
0.3	40	40	5.49 (3.30, 2.19)	0.3201 (0.3187)	5.34 (2.90, 2.44)	0.3200 (0.3189)	4.73 (2.16, 2.57)	0.3205 (0.3197)
	70	70	5.54 (2.89, 2.65)	0.2781 (0.2776)	5.51 (2.80, 2.71)	0.2780 (0.2776)	5.29 (2.55, 2.74)	0.2780 (0.2776)
	120	120	5.19 (2.57, 2.62)	0.2580 (0.2578)	5.25 (2.60, 2.65)	0.2581 (0.2579)	5.08 (2.58, 2.50)	0.2582 (0.2579)
0.5	40	40	4.97 (2.66, 2.31)	0.3074	4.75 (2.44, 2.31)	0.3072	4.30 (2.17, 2.13)	0.3070
	70	70	5.09 (2.42, 2.67)	0.2548	5.04 (2.41, 2.63)	0.2547	4.72 (2.30, 2.42)	0.2548
	120	120	5.23 (2.66, 2.57)	0.2246	5.24 (2.62, 2.62)	0.2246	4.99 (2.68, 2.31)	0.2250
0.7	40	40	5.73 (2.41, 3.32)	0.2568	5.48 (2.49, 2.99)	0.2573	4.76 (2.47, 2.29)	0.2586
	70	70	5.22 (2.28, 2.94)	0.2065	5.14 (2.32, 2.82)	0.2068	4.76 (2.41, 2.35)	0.2078
	120	120	5.63 (2.32, 3.31)	0.1760	5.53 (2.41, 3.12)	0.1762	5.38 (2.73, 2.65)	0.1770
0.9	40	40	6.30 (1.62, 4.68)	0.1517	5.53 (1.76, 3.77)	0.1542	3.83 (2.18, 1.65)	0.1613
	70	70	5.89 (1.58, 4.31)	0.1202	5.33 (1.83, 3.50)	0.1213	4.29 (2.36, 1.93)	0.1250
	120	120	5.64 (1.63, 4.01)	0.1004	5.49 (1.91, 3.58)	0.1011	4.65 (2.36, 2.29)	0.1032
0.95	40	40	9.15 (0.85, 8.30)	0.1020	7.07 (1.04, 6.03)	0.1060	5.16 (1.59, 3.57)	0.1173
	70	70	7.63 (1.18, 6.45)	0.0829	6.48 (1.49, 4.99)	0.0850	4.63 (2.22, 2.41)	0.0907
	120	120	6.82 (1.31, 5.51)	0.0682	5.99 (1.48, 4.51)	0.0708	4.37 (2.30, 2.07)	0.0742
0.5	100	50	5.30 (2.79, 2.51)	0.2337	5.37 (2.72, 2.65)	0.2336	5.03 (2.56, 2.47)	0.2336
0.5	100	80	5.23 (2.68, 2.55)	0.2364	5.17 (2.65, 2.52)	0.2365	5.03 (2.65, 2.38)	0.2367
0.7	70	40	5.62 (2.33, 3.29)	0.2150	5.35 (2.20, 3.15)	0.2149	4.81 (2.17, 2.64)	0.2157
0.7	80	70	5.48 (2.54, 2.94)	0.2009	5.51 (2.63, 2.88)	0.2012	5.21 (2.83, 2.38)	0.2021

Shown are non-coverage percentages (with left and right non-coverage percentages) and mean length of (1) 95% percentile bootstrap c.i., (2) 95% bias-corrected bootstrap c.i., and (3) 95% accelerated bias-corrected (BC$_a$) bootstrap c.i species of sunflowers, Rieseberg (1996) reports that of 220 pairwise comparisons of comigrating fragments only 79% identified loci useful for comparative genetic studies. For RAPD comparable corrections for homoplasy can be envisioned, as we propose here for AFLP.

Software in R (R Development Core Team 2005) for calculation of the proposed estimators is available from the authors.

Acknowledgments We thank Wim Koopman for providing the data from the lettuce study, Herman Adèr for discussing bootstrap procedures, and Alfred Stein for proof reading the manuscript.

Appendix 1: Comparison of bootstrap confidence intervals

We compare three types of bootstrap confidence intervals (c.i.):

- simple percentile c.i.
- bias-corrected percentile c.i.
- accelerated bias-corrected percentile (BC_a) c.i.

These c.i.'s are calculated as described in (Manly 1997), pp. 39–56. For the accelerated bias-corrected percentile c.i.'s calculation of the constant a_{acc} is required. Manly (1997) suggests to approximate a_{acc} by $\sum_{j=1}^{N}(\hat{\Theta}_{.} - \hat{\Theta}_{-j})^3 / [6\{\sum_{j=1}^{N}(\hat{\Theta}_{.} - \hat{\Theta}_{-j})^2\}^{1.5}]$ with $\hat{\Theta}_{-j}$ the partial estimate of the parameter Θ based on all but the jth observation, and

Table 6 Comparison of bootstrap confidence intervals for p_{gs} from a simulation study on D_L^{mod} for a range of values of genetic similarity p_{gs} and expected numbers of fragments m_1 and m_2, 10,000 replicated pairs of AFLP profiles, 1,000 bootstrap resamples, *fld* F_S from *A. thaliana* with $N = 450$

Parameter settings			Results for D_L^{mod}: 95% bootstrap confidence intervals for p_{gs}					
			Percentile bootstrap c.i.		Bias corrected bootstrap c.i.		BC_a c.i.	
p_{gs}	m_1,	m_2	Non-coverage (too low, too high)	Length (trunc)	Non-coverage (too low, too high)	Length (trunc)	Non-coverage (too low, too high)	Length (trunc)
	40	40	6.76 (5.22, 1.54)	0.2527 (0.1339)	6.35 (4.58, 1.77)	0.2544 (0.1394)	5.55 (3.15, 2.40)	0.2605 (0.1524)
	70	70	5.36 (3.30, 2.06)	0.2718 (0.1354)	5.33 (3.12, 2.21)	0.2720 (0.1369)	5.17 (2.68, 2.49)	0.2725 (0.1409)
	120	120	5.39 (2.88, 2.51)	0.2943 (0.1419)	5.60 (3.06, 2.54)	0.2944 (0.1413)	5.55 (2.99, 2.56)	0.2944 (0.1419)
0.1	40	40	5.80 (4.18, 1.62)	0.2913 (0.2378)	5.44 (3.59, 1.85)	0.2922 (0.2419)	4.52 (2.28, 2.24)	0.2957 (0.2510)
	70	70	5.35 (3.12, 2.23)	0.2812 (0.2273)	5.21 (2.88, 2.33)	0.2812 (0.2285)	5.05 (2.50, 2.55)	0.2815 (0.2311)
	120	120	5.21 (2.87, 2.34)	0.2860 (0.2252)	5.22 (2.97, 2.25)	0.2861 (0.2247)	5.10 (2.84, 2.26)	0.2861 (0.2249)
0.3	40	40	5.64 (3.44, 2.20)	0.3212 (0.3197)	5.30 (3.03, 2.27)	0.3210 (0.3197)	4.78 (2.34, 2.44)	0.3213 (0.3204)
	70	70	5.44 (2.87, 2.57)	0.2800 (0.2794)	5.37 (2.69, 2.68)	0.2799 (0.2793)	5.11 (2.43, 2.68)	0.2798 (0.2793)
	120	120	5.36 (2.54, 2.82)	0.2619 (0.2616)	5.26 (2.56, 2.70)	0.2620 (0.2616)	5.14 (2.56, 2.58)	0.2622 (0.2618)
0.5	40	40	5.14 (2.77, 2.37)	0.3078	4.99 (2.58, 2.41)	0.3077	4.38 (2.17, 2.21)	0.3075
	70	70	5.17 (2.50, 2.67)	0.2556	5.04 (2.44, 2.60)	0.2556	4.80 (2.36, 2.44)	0.2558
	120	120	5.44 (2.79, 2.65)	0.2269	5.44 (2.77, 2.67)	0.2271	5.26 (2.83, 2.43)	0.2275
0.7	40	40	5.73 (2.45, 3.28)	0.2567	5.36 (2.45, 2.91)	0.2573	4.67 (2.43, 2.24)	0.2587
	70	70	5.26 (2.26, 3.00)	0.2070	5.05 (2.30, 2.75)	0.2074	4.69 (2.41, 2.28)	0.2085
	120	120	5.67 (2.26, 3.41)	0.1774	5.58 (2.36, 3.22)	0.1777	5.29 (2.51, 2.78)	0.1786
0.9	40	40	6.34 (1.60, 4.74)	0.1518	5.32 (1.77, 3.55)	0.1545	3.78 (2.32, 1.46)	0.1618
	70	70	5.93 (1.63, 4.30)	0.1203	5.39 (1.90, 3.49)	0.1215	4.28 (2.43, 1.85)	0.1253
	120	120	5.78 (1.62, 4.16)	0.1010	5.37 (1.86, 3.51)	0.1017	4.62 (2.44, 2.18)	0.1040
0.95	40	40	8.78 (0.78, 8.00)	0.1020	6.50 (1.03, 5.47)	0.1066	5.26 (1.74, 3.52)	0.1188
	70	70	7.49 (1.12, 6.37)	0.0831	6.33 (1.49, 4.84)	0.0853	4.02 (2.21, 1.81)	0.0914
	120	120	6.96 (1.26, 5.70)	0.0701	5.99 (1.58, 4.41)	0.0713	4.43 (2.34, 2.09)	0.0749
0.5	100	50	5.61 (2.77, 2.84)	0.2342	5.49 (2.70, 2.79)	0.2342	5.19 (2.60, 2.59)	0.2342
0.5	100	80	5.25 (2.66, 2.59)	0.2378	5.19 (2.68, 2.51)	0.2379	5.09 (2.69, 2.40)	0.2381
0.7	70	40	5.75 (2.41, 3.34)	0.2150	5.53 (2.39, 3.14)	0.2151	4.97 (2.37, 2.60)	0.2160
0.7	80	70	5.28 (2.49, 2.79)	0.2017	5.23 (2.56, 2.67)	0.2020	4.94 (2.71, 2.23)	0.2030

Shown are non-coverage percentages (with left and right non-coverage percentages) and mean length of (1) 95% percentile bootstrap c.i., (2) 95% bias-corrected bootstrap c.i., and (3) 95% accelerated bias-corrected (BC_a) bootstrap c.i

$\hat{\Theta}_{.}$ the average of $\hat{\Theta}_{-j}$ ($j = 1,\ldots, N$). In our case the parameter Θ is the fraction of common fragments p_{gs}, estimated by either D_L^{mod} or $D_{\bar{L}}^{\text{mod}}$.

For D_L^{mod} we take a pair of binary scores (y_{1j}, y_{2j}) ($j = 1,\ldots, N$) to be an observation. The constant a_{acc} is calculated by removing observation j from the pair of profiles, rescaling the fragment length distribution, calculating D_L^{mod} from the reduced dataset, and repeating over all band positions ($j = 1,\ldots, N$), resulting in partial estimates $\hat{\Theta}_{-j}$.

For $D_{\bar{L}}^{\text{mod}}$ the information on band lengths is missing, and a pair of profiles can be summarized as a vector of counts (a, b, c, d). The observations are the pairs of binary scores 1–1 (occurring a times), 1–0 (b times), 0–1 (c times), and 0–0 (d times). The partial estimates $\hat{\Theta}_{-j}$ consist of weighted averages [with weights (a, b, c, d)] of $D_{\bar{L}}^{\text{mod}}$ values. We label the weighted averages $\hat{\Theta}^a_{-j}$ (occurring a times), $\hat{\Theta}^b_{-j}$ (b times), $\hat{\Theta}^c_{-j}$ (c times), and $\hat{\Theta}^d_{-j}$ (d times). $\hat{\Theta}^a_{-j}$ is the weighted average of the 4 $D_{\bar{L}}^{\text{mod}}$ values calculated for the profile pairs (a, b, c, d), $(a-1, b+1, c, d), (a-1, b, c+1, d), (a-1, b, c, d+1)$, $\hat{\Theta}^b_{-j}$ is calculated from profile pairs $(a+1, b-1, c, d)$, (a, b, c, d), $(a, b-1, c+1, d)$, $(a, b-1, c, d+1)$, $\hat{\Theta}^c_{-j}$ from profile pairs $(a+1, b, c-1, d), (a, b+1, c-1, d), (a, b, c, d), (a, b, c-1, d+1)$, and $\hat{\Theta}^d_{-j}$ from profile pairs $(a+1, b, c, d-1), (a, b+1, c, d-1), (a, b, c+1, d-1), (a, b, c, d)$.

For the simulation dataset with 10,000 replicates, we calculated 95% bootstrap c.i.'s for $D_{\bar{L}}^{\text{mod}}$, based on a bootstrap resample size of 1000. The results are shown in Table 5. The non-coverage rates for the 95% simple percentile c.i. range from 0.0497 to 0.0915 (average 0.0581), a bit larger than the nominal 0.05. The larger error rates occur for the profiles with smallest expected fragment counts ($m = 40$), and extreme values of p_{gs} ($p_{gs} = 0.0, 0.9, 0.95$). In general the c.i.'s are slightly too narrow. The 95% bias-corrected percentile c.i.'s have better non-coverage rates, ranging from 0.0475 to 0.0707 (average 0.0550). The non-coverage rates of the 95% BC_a c.i.'s range from 0.0383 to 0.0545 (average 0.0486). This last method seems to be a bit too conservative, delivering intervals which are slightly too wide. Over the whole range of p_{gs} values this last method performed best.

For the same simulation data we calculated 95% bootstrap c.i.'s for D_L^{mod} (see Table 6). The non-coverage rates for the simple percentile method range from 0.0514 to 0.0878 (average 0.0584), for the bias-corrected method from 0.0499 to 0.065 (average 0.0548), and for the accelerated bias-corrected method from 0.0378 to 00555 (average 0.0487). Again, the accelerated bias-corrected method performs best with slightly conservative c.i.'s.

Appendix 2

See Table 7 for a list of used symbols.

Table 7 Overview on symbols

Symbol	Description	Type
N	Number of observable band lengths, derived from scoring range; e.g. 450 if scoring range is 51–500	Constant
i	Index of band length ($i = 1,\ldots, N$)	Index
j	Index of lane number or genotype number ($j = 1, 2$)	Index
Π_j	Population of fragments after restriction, eligible for visualization, for genotype j	Population
M_j	Number of fragments of Π_j	Parameter
p_i	Probability that a fragment randomly drawn from Π has length i	Constant
fld	Fragment length distribution = $(p_1,\ldots, p_N)$	Constant
F_S	*Fld* from in silico AFLP for *A. thaliana*, see Gort et al. (2006)	Constant
π	Probability of a fragment in Π to be sampled	Parameter
m_j	Expected number of fragments in jth lane = πM_j, proportional to M_j	Parameter
k_j	Number of fragments in lane j; distributed as Poisson (m_j)	Stochastic
y_{ij}	Binary score for absence/presence of a band of length i in lane j	Stochastic
n_j	Number of bands in jth lane = $\sum_{i=1}^N y_{ij}$	Stochastic
Π_a	Population of common fragments; $\Pi_1 \cap \Pi_2$	Population
Π_b	Population of fragments unique to genotype 1; $\Pi_1 \cap \bar{\Pi}_2$	Population
Π_c	Population of fragments unique to genotype 2; $\bar{\Pi}_1 \cap \Pi_2$	Population
F_j	Fraction of common fragments in jth population = M_a/M_j	Parameter
p_{gs}	Pairwise genetic similarity for AFLP = $\frac{M_1}{M_1+M_2}F_1 + \frac{M_2}{M_1+M_2}F_2$	Parameter
a	Number of shared bands in the two profiles = $\sum_{i=1}^N y_{i1}y_{i2}$	Stochastic
b	Number of bands in the first profile, which are absent in the second = $\sum_{i=1}^N y_{i1}(1-y_{i2})$	Stochastic
c	Number of bands in the second profile, which are absent in the first = $\sum_{i=1}^N (1-y_{i1})y_{i2}$	Stochastic

Table 7 continued

Symbol	Description	Type
d	Number of empty positions in both profiles $= \sum_{i=1}^{N}(1-y_{i1})(1-y_{i2})$	Stochastic
D	Dice coefficient $= 2a/(2a+b+c)$	Stochastic
J	Jaccard coefficient $= a/(a+b+c)$	Stochastic
P_i	Probability of a band of length i, given m fragments $= 1-(1-p_i)^m$	Parameter
$\hat{m}_{\bar{L}}$	Estimator of m without band length information; see Gort et al. (2006)	Stochastic
$\hat{m}_L$	Estimator of m with band length information, based on g.l.m.	Stochastic
$\hat{m}$	Estimator of m with band length information, based on m.l.	Stochastic
$D_{\bar{L}}^{\mathrm{mod}}$	Modified Dice coefficient, without band length info $= 2\hat{m}_{\bar{L}a}/(2\hat{m}_{\bar{L}a}+\hat{m}_{\bar{L}b}+\hat{m}_{\bar{L}c})$	Stochastic
D_L^{mod}	Modified Dice coefficient, with band length info $= 2\hat{m}_{La}/(2\hat{m}_{La}+\hat{m}_{Lb}+\hat{m}_{Lc})$	Stochastic
$J_{\bar{L}}^{\mathrm{mod}}$	Modified Jaccard coefficient, without band length info $= \hat{m}_{\bar{L}a}/(\hat{m}_{\bar{L}a}+\hat{m}_{\bar{L}b}+\hat{m}_{\bar{L}c})$	Stochastic
J_L^{mod}	Modified Jaccard coefficient, with band length info $= \hat{m}_{La}/(\hat{m}_{La}+\hat{m}_{Lb}+\hat{m}_{Lc})$	Stochastic
D_1^{mle}	Modified Dice coefficient based on m.l. estimation of $m = 2\hat{m}_a/(2\hat{m}_a+\hat{m}_b+\hat{m}_c)$	Stochastic
D_2^{mle}	Direct m.l. estimator of p_{gs}	Stochastic
D_3^{mle}	Backtransformed estimator of p_{gs} (using m.l. estimation of logit(p_{gs}))	Stochastic

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

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

Caballero A, Quesada H, Rolán-Alvarez E (2008) Impact of amplified fragment length polymorphism size homoplasy on the estimation of population genetic diversity and the detection of selective loci. Genetics 179:539–554

Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302

DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. Stat Sci 11:189–228

Drossou A, Katsiotis A, Leggett JM, Loukas M, Tsakas S (2004) Genome and species relationships in genus Avena based on RAPD and AFLP molecular markers. Theor Appl Genet 109:48–54

Duim B, Vandamme PAR, Rigter A, Laevens S, Dijkstra JR, Wagenaar JA (2001) Differentiation of *Campylobacter* species by AFLP fingerprinting. Microbiology 147:2729–2737

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall, New York

Foulley JL, van Schriek MGM, Alderson L, Amigues Y, Bagga M, Boscher MY, Brugmans B, Cardellino R, Davoli R, Delgado JV, Fimland E, Gandini GC, Glodek P, Groenen MAM, Hammond K, Harlizius B, Heuven H, Joosten R, Martinez AM, Matassino D, Meyer JN, Peleman J, Ramos AM, Rattink AP, Russo V, Siggens KW, Vega-Pla JL, Ollivier L (2006) Genetic diversity analysis using lowly polymorphic dominant markers: the example of AFLP in pigs. J Hered 97:244–252

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

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, Meudth HM (2008) Optimizing automated AFLP scoring parameters to improve phylogenetic resolution. Syst Biol 57:347–366

Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bulletin de la Société Vaudoise des Sciences Naturelles. 44:223–270

Jansen J, van Hintum T (2007) Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. Theor Appl Genet 114:421–428

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

Koopman WJM, Zevenbergen M, Van den Berg R (2001) Species relationships in *Lactuca* s.l. (Lactuceae, Asteraceae) inferred from AFLP fingerprints. Am J Bot 88:1881–1887

Koopman WJM, Hadam J, Doležel J (2002) Evolution of DNA content and base composition in *Lactuca* (Asteraceae) and related genera. In: Zooming in on the lettuce genome (Ph. D. thesis W.J.M. Koopman). Wageningen University, Wageningen, The Netherlands

Krauss SL (2000) Accurate gene diversity estimates from amplified fragment length polymorphism (AFLP) markers. Mol Ecol 9:1241–1245

Manly BFJ (1997) Randomization, bootstrap and Monte Carlo methods in biology. Chapman & Hall, London

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

Mebrate SA, Dehne HW, Pillen K, Oerke EC (2006) Molecular diversity in *Puccinia triticina* isolates from Ethiopia and Germany. J Phytopathol 154:701–710

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

Nei M, Li WH (1979) Mathematical-model for studying genetic-variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

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

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

Reif JL, Melchinger AE, Frisch M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci 45:1–7

Rieseberg LH (1996) Homology among RAPD fragments in interspecific comparisons. Mol Ecol 5:99–105

Robinson JP, Harris SA (1999) Amplified fragment length polymorphisms and microsatellites: a phylogenetic perspective. In: Which DNA marker for which purpose? Chap. 12. Final Compendium of the Research Project Development, optimization and validation of molecular tools for assessment of biodiversity in forest trees. European Union DGXII Biotechnology FW IV Research Programme Molecular Tools for Biodiversity. URL http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm

Silvey SD (1975) Statistical inference. Chapman & Hall, London

Sneath PHA, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco

Tams SH, Melchinger AE, Bauer E (2005) Genetic similarity among European winter triticale elite germplasms assessed with AFLP and comparisons with SSR and pedigree data. Plant Breed 124:154–160

van Eeuwijk FA, Law JR (2004) Statistical aspects of essential derivation, with illustrations based on lettuce and barley. Euphytica 137:129–137

van Berloo R, Zhuy A, Ursem R, Verbakel H, Gort G, Van Eeuwijk FA (2008) Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes. Theor Appl Genet 117:89–101

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

Venzon DJ, Moolgavkar SH (1988) A method for computing profile-likelihood-based confidence intervals. Appl Stat 37:87–94

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

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity arrays technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci USA 101:9915–9920

Wong A, Forbes MR, Smith ML (2001) Characterization of AFLP markers in damselflies: prevalence of codominant markers and implications for population genetic applications. Genome 44:677–684

Zhong D, Menge DM, Temu EA, Chen H, Yan G (2006) Amplified fragment length polymorphism mapping of quantitative trait loci for malaria parasite susceptibility in the yellow fever mosquito *Aedes aegypti*. Genetics 173:1337–1345