DOI: 10.1534/genetics.103.015693

Significance Tests and Weighted Values for AFLP Similarities, Based on Arabidopsis *in Silico* AFLP Fragment Length Distributions

Wim J. M. Koopman*,[1] and Gerrit Gort†

**Nationaal Herbarium Nederland—Wageningen Branch, Biosystematics Group, Wageningen University, 6703 BL Wageningen, The Netherlands and †Biometris, Wageningen University and Research Centre, 6700 AA, Wageningen, The Netherlands*

Manuscript received March 26, 2003
Accepted for publication April 29, 2004

ABSTRACT

Many AFLP studies include relatively unrelated genotypes that contribute noise to data sets instead of signal. We developed: (1) estimates of expected AFLP similarities between unrelated genotypes, (2) significance tests for AFLP similarities, enabling the detection of unrelated genotypes, and (3) weighted similarity coefficients, including band position information. Detection of unrelated genotypes and use of weighted similarity coefficients will make the analysis of AFLP data sets more informative and more reliable. Test statistics and weighted coefficients were developed for total numbers of shared bands and for Dice, Jaccard, Nei and Li, and simple matching (dis)similarity coefficients. Theoretical and *in silico* AFLP fragment length distributions (FLDs) were examined as a basis for the tests. The *in silico* AFLP FLD based on the *Arabidopsis thaliana* genome sequence was the most appropriate for angiosperms. The $G + C$ content of the selective nucleotides in the *in silico* AFLP procedure significantly influenced the FLD. Therefore, separate test statistics were calculated for AFLP procedures with high, average, and low $G + C$ contents in the selective nucleotides. The test statistics are generally applicable for angiosperms with a $G + C$ content of ~35–40%, but represent conservative estimates for genotypes with higher $G + C$ contents. For the latter, test statistics based on a rice genome sequence are more appropriate.

AFLP is a DNA fingerprinting technique developed by Keygene N.V. (Vos *et al.* 1995). The technique consist of four steps: (1) digestion of DNA with two restriction enzymes, (2) ligation of double-stranded oligonucleotide adapters to the restriction fragments, (3) selective PCR amplification of the ligated fragments with specific PCR primers that have selective nucleotides at their 3′ end, and (4) separation of the amplified fragments on a denaturing polyacrylamide gel. On this gel, the fragments are separated by their length. Inclusion of a base-pair ladder enables determination of the exact length of each fragment.

In recent years, AFLPs have become a popular tool for relationship studies (Mueller and LaReesa Wolfenbarger 1999). In these studies, the AFLPs are scored as dominant anonymous markers. Dominant scoring of AFLPs means that each fragment is scored as either present or absent and that the fragments are assumed to occur independently of each other. Scoring as anonymous markers means that the fragments are recognized only by their length, while their sequence is unknown. Fragments of the same length, which are comigrating on a gel, are assumed to be identical. The fraction of fragments comigrating across genotypes, expressed in some way by a similarity or dissimilarity coefficient, is used as a measure for genetic or phenetic relationship. Various coefficients have been developed to quantify (dis)similarity, mainly differing in the weighting of comigrating relative to noncomigrating fragments (see, *e.g.*, Nei and Li 1979; Rohlf 1993).

The assumption that all comigrating fragments are identical is an oversimplification of the actual situation (Vekemans *et al.* 2002). In reality, a certain fraction of fragments will be comigrating by chance only, while having distinct sequences. Because these fragments will be scored as identical, their presence leads to an overestimation of the similarity among genotypes. The presence of nonidentical fragments comigrating across genotypes was demonstrated in actual data sets of *Solanum tuberosum* (Rouppe van der Voort *et al.* 1997), Carduinae thistles (O'Hanlon and Peakall 2000), and Hordeum species (El-Rabey *et al.* 2002). The presence of nonidentical fragments comigrating within genotypes was demonstrated in Beta (Hansen *et al.* 1999) and *Glycine max* (Meksem *et al.* 2001). The proportion of comigrating nonidentical fragments ranged from at least 10% within genotypes or among closely related genotypes (Rouppe van der Voort *et al.* 1997; Hansen *et al.* 1999; Meksem *et al.* 2001) to 100% for pairs of genotypes from more distantly related taxa (O'Hanlon and Peakall 2000). Given the proportions of comigrating nonidentical bands, a serious overestimation of pairwise similarities among genotypes can be expected. Indeed, Karp *et al.* (1996) noted that the occurrence of

[1] *Corresponding author:* Plant Research International B.V., 6700 AA, Wageningen, The Netherlands. E-mail: wim.koopman@wur.nl

nonidentical comigrating AFLP fragments may pose serious problems for the application of AFLPs in relationship studies, but the issue was largely ignored in literature thereafter.

In this study, we quantify the occurrence of nonidentical comigrating AFLP fragments for AFLP procedures with restriction enzymes *Eco*RI/*Mse*I. The estimates are used to (1) determine the expected numbers of comigrating nonidentical bands and (2) develop significance tests for AFLP similarities. As a basis for the significance tests we determine and evaluate theoretical AFLP fragment length distributions based on Innan *et al.* (1999) and *in silico* AFLP fragment length distributions (FLDs) based on the complete *Arabidopsis thaliana* (L.) Heynh. genome sequence (Arabidopsis Genome Initiative 2000). Using the *A. thaliana* (hereafter, Arabidopsis) FLD, we estimate the probability distribution of the number of nonidentical AFLP bands comigrating across genotypes. From this distribution, we determine expectations and 95 and 99% critical values for band numbers and (dis)similarity coefficients Dice, Jaccard, Nei and Li, and simple matching (Nei and Li 1979; Rohlf 1993). The critical values can be used to test the significance of a given pairwise similarity among angiosperm genotypes. If desired, genotypes that do not contribute significant relationship information can be removed from a data set. Determination of the expected numbers of comigrating nonidentical bands also yielded information on the underlying band length distribution probabilities. However, the usual similarities calculated using the Dice, Jaccard, Nei and Li, and simple matching coefficients ignore this information, assuming identical probabilities for all bands. As an alternative, we propose similarity coefficients that weight the AFLP bands according to their band length distribution probabilities. It is expected that the use of the significance tests and weighted similarities will make the analysis of AFLP data sets more informative and more reliable.

METHODS AND RESULTS

General strategy: The number of nonidentical AFLP bands comigrating across genotypes depends on the number of bands scored for each genotype, the number of possible band lengths for the genotypes (*i.e.*, the number of discrete band positions possible within a selected scoring range), and the length distribution of the AFLP fragments. Note that one AFLP band may contain multiple fragments (discussed later). In empirical data sets, the number of possible band positions and the number of bands for each genotype are known; only the FLD remains to be determined. The distribution can be obtained in several ways, *e.g.*, (1) derived from AFLP band data in empirical data sets, (2) calculated using theoretical FLDs, and (3) determined *in silico*, if representative genome sequence data (preferably entire genomes) are available.

The use of empirical data involves the risk of introducing methodological error into the calculations resulting from the AFLP procedure itself. Such errors may include, *e.g.*, biases in fragment amplification or in scoring of bands. Theoretically derived or *in silico*-generated FLDs do not have this drawback.

Theoretical distributions may be preferred over *in silico* distributions, because they are exactly formulated, using explicit assumptions and parameter settings. In this article, we examine the length distribution for AFLP fragments proposed by Innan *et al.* (1999) as a theoretical basis on which to estimate the proportion of nonidentical bands comigrating across genotypes. To our knowledge, no alternative AFLP FLD has been proposed yet.

Use of *in silico* AFLP FLDs has the drawback that the distribution itself has to be estimated from the available genome data. Therefore, it is inherently subject to uncertainty because of estimation error and limited by the availability and representativeness of the genome data. However, *in silico* AFLP data also have two major advantages. First, the AFLP fragments represent an actual genome. Thus, their distribution is not subject to assumptions that underlie theoretical models. Second, when the procedure is performed properly, no fragments will be lost due to methodological errors, and all possible fragments will be represented in the AFLP data set. Here, we examine an *in silico* FLD based on the genome sequence of the model plant Arabidopsis as an alternative to the theoretical distribution of Innan *et al.* (1999). All statistical procedures were performed in SAS Release 8.00 (SAS Institute, Cary, NC).

Theoretical AFLP fragment length distributions: Innan *et al.* (1999) describe AFLP FLDs for *Eco*RI and *Mse*I restriction enzymes under the assumption of (1) a random nucleotide sequence under the Jukes and Cantor model [equal base frequencies $C = A = T = G = 0.25$, and all substitutions equally likely (Jukes and Cantor 1969)]; (2) nucleotide changes as sole cause of changes in DNA sequence; and (3) a haploid genome. They showed that both *Eco*RI/*Eco*RI and *Eco*RI/*Mse*I fragments follow the same truncated geometric distribution $G(L) = ((1 - A)A^{L-L_{\min}})/(1 - A^{L_{\max}-L_{\min}+1})$, in which L is the length of the AFLP fragments, $L_{\min}$ and $L_{\max}$ are the minimum and maximum possible lengths of the fragments considered, and $A = (1 -$ probability of formation of new *Eco*RI site$)(1 -$ probability of formation of new *Mse*I site). The probability of formation of a restriction site equals the multiplied relative frequencies of the individual nucleotides required for such a site (*GAATTC* for *Eco*RI, *TTAA* for *Mse*I). Under the assumption of equal frequencies of occurrence for all four nucleotides as made by Innan *et al.* (1999), $A = (1 - 0.25^6)(1 - 0.25^4)$.

To examine the influence of nucleotide frequencies on the AFLP FLD, we calculated distributions for various ratios of $A + T$ *vs.* $G + C$. A literature survey revealed

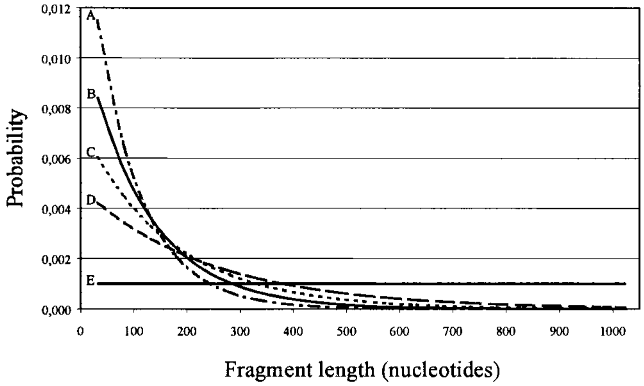

Figure 1.—Theoretical AFLP FLDs based on Innan *et al.* (1999) for a genome with 35% $G + C$ (A), 40% $G + C$ (B), 45% $G + C$ (C), and 50% $G + C$ (D), respectively. The uniform distribution (E; equal probability for all fragments) is given as a reference.

that the $G + C$ contents of the majority of plants ranged between 35 and 50% (see, *e.g.*, Marie and Brown 1993; Barow and Meister 2002). However, various plant groups showed different $G + C$ contents. The average $G + C$ content was 37% for gymnosperms, 40% for dicotyledons, 41% for ferns, 44% for monocotyledons, and 45% for algae. *Viscum album* possibly occupies a special position with only 30% $G + C$ (Nagl and Stein 1989), although Marie and Brown (1993) reported 39% $G + C$. We covered the $G + C$ range by calculating separate AFLP FLDs for 35, 40, 45, and 50% $G + C$. The nucleotide frequencies of A in the formula of Innan *et al.* (1999) were adjusted accordingly, with equal splitting of percentages over $A + T$ and $G + C$ nucleotides. For easy comparison with empirical data sets, all fragment and band lengths that are reported in this article include adapter sequences.

Figure 1 depicts the AFLP FLDs for 35–50% $G + C$. The distributions show that the probability that a fragment will occur decreases with increasing fragment length for all $G + C$ contents. The shape of the distribution is also influenced by the base composition: low $G + C$ contents yield relatively high frequencies of smaller fragments, while high $G + C$ contents yield relatively high frequencies of longer fragments. The uniform distribution (all fragment lengths equally likely) is given as a reference.

Arabidopsis *in silico* AFLP fragment length distributions: Sequence data of the entire Arabidopsis genome sequence were obtained from The Institute for Genomic Research through the web site at http://www.tigr.org. The Arabidopsis *in silico* AFLP was performed using the restriction enzyme sequences of *Eco*RI/*Mse*I without any selective nucleotides. The probability distribution of the fragment lengths was estimated by fitting a cubic smoothing spline and rescaling properly, using SAS PROC IML. The smoothing parameter of the spline (200.000) was chosen by eye. The more objective approach of cross-validation (SAS PROC INSIGHT) resulted in an unsatisfactory smoothing level and a spline oscillating around the one chosen by eye. The smoothing spline and the relative frequency distribution of the *in silico* AFLP fragments are depicted in Figure 2. Fragment lengths range from 32 to 1024 bp.

To compare the *in silico* AFLP FLD with the theoretical distribution of Innan *et al.* (1999), we calculated a theoretical distribution using the nucleotide frequencies from the Arabidopsis genome sequence ($G = C = 0.18$ and $A = T = 0.32$ for all five chromosomes). Figure 2 shows a clear difference between the theoretical and the *in silico* FLD. Compared to the theoretical distribution, the *in silico* distribution shows a lack of smaller bands (<179 bp) and an excess of larger bands (>179 bp). The difference may originate in the nucleotide sequence model employed by Innan *et al.* (1999), which was probably too simple to adequately describe the Arabidopsis *in silico* FLD (see discussion). Given the limitations of the theoretical model and the fact that, in contrast, the Arabidopsis *in silico* FLD reflects an actual genome sequence, we consider the Arabidopsis distribution to be the more accurate basis for our significance tests for AFLP similarities.

The *in silico* AFLP FLD was generated without selective nucleotides to obtain the highest possible number of AFLP fragments. In practice, however, selective nucleotides are always employed in AFLP procedures on plants. To test the influence of selective nucleotides on the AFLP FLD, we performed additional *in silico* AFLP runs with three +1/+1 selective nucleotide combinations: *A*/*C* (the most commonly used single-nucleotide combination), *T*/*A* (the nucleotides with the highest frequency in the Arabidopsis genome), and *C*/*G* (the nucleotides with the lowest frequency in the Arabidopsis genome). A two-sample Kolmogorov-Smirnov test (SAS PROC NPAR1WAY) showed a significant influence of *T*/*A* ($P = 0.002$) and *C*/*G* ($P = 0.001$) selective nucleotides on the FLD. The distribution for selective nucleotides *A*/*C* did not differ significantly from that without selective nucleotides ($P = 0.62$). Figure 2 illustrates the influence of selective nucleotides on the *in silico* AFLP FLD. The use of *T*/*A* selective nucleotides results in an overrepresentation of shorter fragments (<107 bp) and an underrepresentation of longer fragments (>107 bp). The use of *C*/*G* selective nucleotides results in an overrepresentation of longer fragments (>107 bp) and an underrepresentation of shorter fragments (<107 bp). The difference indicates that selection of AFLP fragments using selective nucleotides is not a random process (see discussion).

Each fragment in an AFLP profile contains a discrete number of nucleotides. If properly measured, the length of a fragment equals this number of nucleotides. Given the discrete nature of the AFLP fragment lengths, the AFLP FLDs are discrete distributions. In Figures 2 and 4, however, the AFLP FLDs appear as continuous

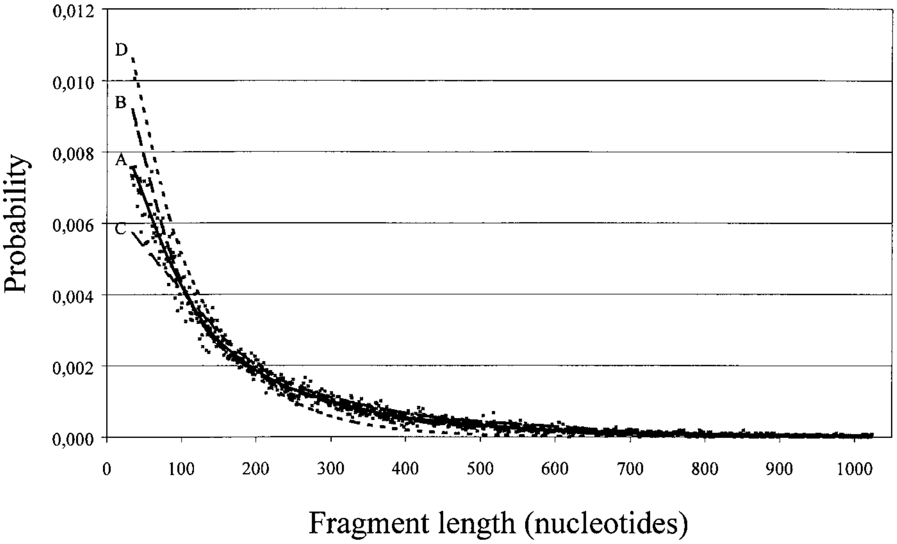

Figure 2.—Relative frequency distribution of fragments resulting from *in silico* AFLP on the Arabidopsis genome sequence without selective nucleotides (frequencies for each length class are denoted by dots). (A) Smoothed FLD resulting from *in silico* AFLP on the Arabidopsis genome sequence without selective nucleotides (note that this distribution is not significantly different from a distribution with *A/C* selective nucleotides). (B) Smoothed FLD resulting from *in silico* AFLP on the Arabidopsis genome sequence with *T/A* selective nucleotides. (C) Smoothed FLD resulting from *in silico* AFLP on the Arabidopsis genome sequence with *C/G* selective nucleotides. (D) Theoretical AFLP FLD based on Innan *et al.* (1999) for a genome with 36% *G* + *C*. Fragment lengths range from 32 to 1024 bp.

distributions, because the large number of possible lengths makes it impossible to visualize the actual discreteness. For the *in silico* AFLPs without selective nucleotides, Figures 2 and 4 show both the smoothed discrete FLDs (line A in Figure 2; lines A and B in Figure 4) and the nonsmoothed discrete FLDs (probability in each length class depicted by a dot). All statistical procedures in this study are based on the discrete smoothed distributions. As a consequence, band lengths used as input for the statistical tests developed in our study should be discrete (*i.e.*, integer) values.

AFLP fragments and AFLP bands: Similarities in AFLP patterns result from fragments that are comigrating across genotypes, and two types of such fragments can be distinguished: first, fragments that share the same sequence and originate from the same loci (comigrating identical fragments; these fragments reflect the genetic similarity among genotypes); and second, fragments having different sequences, originating from different loci (comigrating nonidentical fragments; these fragments comigrate by chance only, and do not reflect genetic similarity). Genotypes that are too distantly related for the AFLP technique to detect any relationship information (called "unrelated" hereafter) share only the second type of fragments. Therefore, an estimate of the number of nonidentical fragments comigrating across genotypes is an estimate of the lower boundary for fragment similarity to indicate relationship. We use this number to derive test statistics for significance tests on pairwise AFLP similarities between genotypes.

In an ideal situation, each AFLP band consists of only one AFLP fragment, enabling a one-to-one translation of AFLP fragments into AFLP bands. In that case, test statistics for significance tests can be based directly on the numbers of nonidentical fragments comigrating across genotypes. In practice, however, an AFLP band often contains multiple fragments that are comigrating within the same genotype. As a result, identical bands comigrating across genotypes may contain both identical and nonidentical fragments, while nonidentical bands comigrating across genotypes each may contain multiple nonidentical fragments. The phenomenon of nonidentical comigrating fragments (both within and across genotypes) is known as size homoplasy (Vekemans *et al.* 2002). In most relationship studies this size homoplasy is ignored, and only the presence or absence of AFLP bands is recorded. As a result, the similarities calculated in these studies are based on AFLP band similarities rather than on AFLP fragment similarities. For significance tests to be readily applicable in such relationship studies, the test statistics should be derived from the numbers and positions of nonidentical bands comigrating across genotypes. To account for the size homoplasy, however, information on the numbers and positions of nonidentical fragments comigrating across genotypes should be included as well. We constructed a series of significance tests that meet both demands. To our knowledge, there is no straightforward analytical procedure to calculate the relationship between the numbers of AFLP fragments and numbers of AFLP bands. Therefore, we estimated this relationship using Monte Carlo simulations.

Significance tests for pairwise AFLP band similarities: The significance tests for pairwise AFLP band similarities were developed in three steps. In the first step, probability distributions, *P*, of the numbers of nonidentical bands comigrating across genotypes were determined. In the second step, from *P* the expectation, standard deviation, and approximate critical values (95 and 99%) of numbers of nonidentical bands comigrating across genotypes were determined. In the third step, the same quantities were determined for four widely employed (dis)similarity coefficients.

1. For each pairwise comparison, two independent AFLP band patterns were generated with the appro-

priate numbers of bands (*e.g.*, 50 and 60). The band patterns were generated by randomly drawing fragments from the smoothed Arabidopsis AFLP FLD. Note that the fragments are drawn only from the part of the Arabidopsis AFLP FLD corresponding to the scoring range of interest (*e.g.*, 50–500 bp). The numbers of fragments needed for each band pattern were often higher than the numbers of bands in the patterns, because some of the fragments ended up in the same bands. The difference between the numbers of fragments and the numbers of bands indicates the amount of size homoplasy in the band pattern (see also *Nonidentical AFLP fragments comigrating within genotypes*).

To determine the number of fragments to be drawn from the AFLP FLD in an unbiased way, we repeatedly drew a fragment count from a uniform distribution. Next, a number of fragments equal to the fragment count was drawn from the smoothed Arabidopsis FLD, and the resulting number of AFLP bands was determined. The procedure was repeated until the appropriate numbers of bands (*e.g.*, 50 and 60) were reached in both AFLP patterns. For these numbers of bands, the number of bands comigrating across both AFLP patterns was determined and recorded. The entire procedure was repeated 1,000,000 times, and the probability distribution *P* was estimated from the scores of all 1,000,000 replications.

2. In the second step, expected numbers of nonidentical bands comigrating across genotypes (*i.e.*, expected numbers of bands comigrating by chance), standard deviation, and approximate critical values (95 and 99%) were determined from the probability distribution *P*. Because the variables under study are discrete, exact 95 and 99% critical values could not be calculated. Instead, approximate values were determined by interpolation.
3. In most relationship studies, similarity among genotypes is reported using (dis)similarity coefficients rather than numbers of comigrating bands. These coefficients somehow express the proportion of comigrating relative to noncomigrating bands. A literature survey showed that the majority of studies employed Dice similarity (Dice 1945) or Nei and Li distance (Nei and Li 1979), while Jaccard (Jaccard 1908) and simple matching (Sokal and Sneath 1963) similarity are also widely employed. For a given pair of genotypes, let $x_i = 0$ when no AFLP band is present at position *i* in genotype 1, and $x_i = 1$ when an AFLP band is present at position *i* in genotype 1. Likewise, $y_i = 0$ or 1 for genotype 2. For a scoring range 1–*N*, let $s_i = 1$ when a certain band position is scored a data set and $s_i = 0$ when a band position is not scored. Let $a = \sum_{i=1}^{N} x_i y_i s_i$, $b = \sum_{i=1}^{N} x_i(1 - y_i)s_i$, $c = \sum_{i=1}^{N}(1 - x_i)y_i s_i$, and $d = \sum_{i=1}^{N}(1 - x_i)(1 - y_i)s_i$. Then Dice $= 2a/(2a + b + c)$, Jaccard $= a/(a + b + c)$, and simple matching $= (a + d)/(a + b + c + d)$. Nei and Li $= (1 - \text{Dice})$. To make our tests readily applicable in relationship studies employing the above coefficients, we used the numbers of nonidentical bands comigrating across genotypes to get (dis)similarity values. The recalculations involved two steps. First, probability distributions for all four coefficients were calculated, on the basis of the probability distribution of the number of comigrating bands, *P*. Next, expected values and approximate critical values (95 and 99%) were determined from these distributions as described previously.

The entire procedure has been incorporated in the computer program AFLSIM, which can be downloaded from http://www.dpw.wur.nl/biosys/AFLSIM_UK.html. The program can be used to test the significance of AFLP similarities in empirical data sets with scoring ranges between 34 and 1024 bp (related to the limits of the Arabidopsis AFLP FLD). The minimum number of AFLP bands per genotype should be 1, and the maximum equals half the number of band positions available within the employed scoring range. Band lengths should be input as discrete (*i.e.*, integer) values. As an example, Figure 3 and Table 1 show results for the widely employed scoring range 50–500 bp and an AFLP procedure with *A*/*C* selective nucleotides. Figure 3 shows the relationship between the number of bands scored in each of two genotypes and the expected number of bands shared. Table 1 gives an overview of the test statistics. The expected (dis)similarities in the table indicate the level of (dis)similarity expected in unrelated genotypes. Pairwise (dis)similarities exceeding the critical values indicate significant phenetic or genetic similarity.

For the calculations in Table 1, we assumed that all band positions available in the scoring range were present in the data set. As a result, a relatively large proportion of the band positions showed 0/0 matches (*i.e.*, no band present in either of the genotypes compared). Because 0/0 matches are counted as similarity in the simple matching coefficient, this causes a relatively high minimum simple matching value (Table 1, bottom, column 10). The number of 0/0 matches does not influence the Dice, Nei and Li, and Jaccard similarity. Consequently, the theoretical minimum value of these coefficients is always 0, regardless of the number of 0/0 matches in the data set.

The maximum possible (dis)similarity values (given the observed band numbers; see Table 1) illustrate an often overlooked peculiarity of Dice, Jaccard, Nei and Li, and simple matching pairwise (dis)similarities: they can be unity (or 0 in the case of Nei and Li distance) only when AFLP band numbers in both genotypes are identical. Table 1 shows that the maximum possible similarity rapidly decreases with increasing difference in band number between genotypes. Comparison with the critical values corresponding to the unequal band

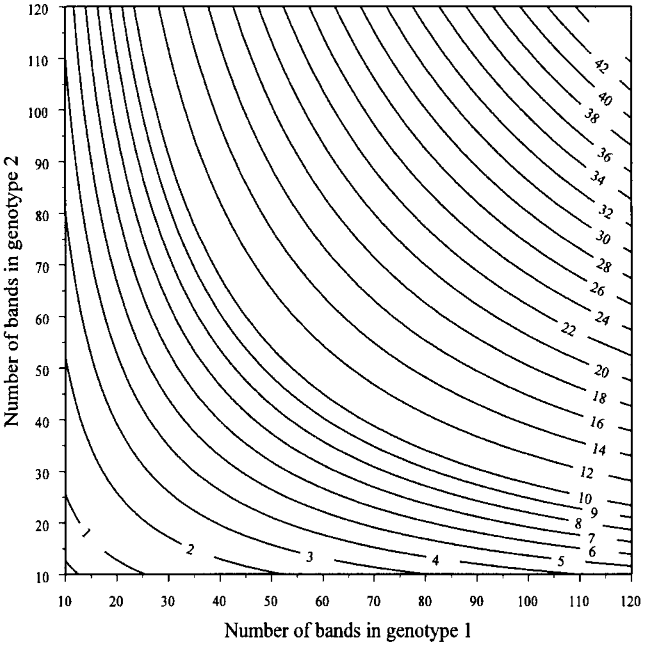

Figure 3.—Relationship between number of bands scored in each of two genotypes and the expected number of bands shared. The lines depict whole numbers of expected shared bands; the actual numbers are inserted in the lines at the bottom and the right side of the plot. The plot corresponds to a scoring range of 50–500 bp and an AFLP procedure with *A/C* selective nucleotides.

numbers shows that such (dis)similarities, although low, may still be significant.

Nonidentical AFLP fragments comigrating within genotypes: When simulating band patterns for the probability distribution *P*, we were surprised by the high amount of size homoplasy. The number of bands containing multiple fragments was much higher than we intuitively anticipated. However, the phenomenon that a co-occurrence of events (in this case the appearance of two AFLP fragments of equal length) is more likely than intuitively expected is well known in statistics and commonly referred to as the birthday paradox. The paradox is often summarized as follows: in a group of only 23 persons, the probability of at least one coinciding birthday, assuming uniformly distributed birthdays over all 365 days of the year, is already >0.5.

Translated to AFLP patterns for a scoring range of, *e.g.*, 50–500 bp (451 positions), this means that only 26 fragments are needed to have a probability >0.5 that at least one AFLP band contains multiple fragments. In reality, however, the probability distribution of fragment lengths is highly skewed instead of uniform (Figure 2), rendering even higher probabilities of fragments with identical lengths (Munford 1977).

Analogous to the situation for nonidentical AFLP bands comigrating across genotypes, the number of nonidentical AFLP fragments comigrating within a genotype (*i.e.*, the amount of size homoplasy) depends on the number of bands scored, the number of discrete band positions available within the scoring range, and the AFLP FLD. Table 2 illustrates the size homoplasy for a wide series of scoring ranges and band numbers. The table shows that the amount of size homoplasy increases with increasing numbers of bands and with decreasing scoring range. In empirical data sets, the occurrence of multiple fragments in AFLP bands has already been demonstrated for Beta and *G. max* (Hansen *et al.* 1999; Meksem *et al.* 2001).

Weighted similarity coefficients including band position information: In the previous sections, a procedure was developed to test the significance of AFLP-based similarities. The procedure can be used to test similarities that were calculated according to various well-known similarity coefficients. The relationship between band length and band presence is incorporated in the tests using the Arabidopsis AFLP FLD. However, this relationship is not accounted for in the similarity coefficients themselves, since all bands are equally weighted in the existing coefficients.

To make the existing similarity coefficients more informative, we propose an adjustment of these coefficients by weighting the bands with the inverse probabilities of their occurrence in an AFLP profile. The rationale behind this is that long bands have a smaller probability of occurring than short bands do, and therefore they have a larger probability of contributing reliable information to a data set. Consequently, long bands should contribute more to the overall similarity values. A proper weighting scheme can be derived from the Arabidopsis AFLP FLD. In the section on Arabidopsis *in silico* AFLP FLDs, we demonstrated that the Arabidopsis AFLP FLD is a reliable basis for describing the probabilities of occurrence of AFLP fragments and hence of AFLP bands. Therefore, the inverse probabilities from the Arabidopsis AFLP FLD are the logical basis for constructing weighted similarity coefficients.

The weighted coefficients are constructed in two steps, analogous to the construction of the unweighted coefficients. In the first step, weighted similarities are calculated for numbers of bands shared between two genotypes (a_w), for numbers of bands unique to one of the genotypes (b_w and c_w), and for band positions that are not occupied in either of the genotypes (d_w). Again, for a given pair of genotypes, let $x_i = 0$ when no AFLP band is present at position *i* in genotype 1, and $x_i = 1$ when an AFLP band is present at position *i* in genotype 1. Likewise, $y_i = 0$ or 1 for genotype 2. For a scoring range 1–*N*, let $s_i = 1$ when a certain band position is scored a data set, and $s_i = 0$ when a band position is not scored. Then, $a_w = N \sum_{i=1}^{N} w_{ai} x_i y_i s_i / \sum_{i=1}^{N} w_{ai}$, $b_w = N \sum_{i=1}^{N} w_{bi} x_i (1 - y_i) s_i / \sum_{i=1}^{N} w_{bi}$, $c_w = N \sum_{i=1}^{N} w_{ci} (1 - x_i) y_i s_i / \sum_{i=1}^{N} w_{ci}$, and $d_w = N \sum_{i=1}^{N} w_{di} (1 - x_i)(1 - y_i) s_i / \sum_{i=1}^{N} w_{di}$; with inverse weights $w_{ai}^{-1} = p_i q_i$, $w_{bi}^{-1} = p_i(1 - q_i)$, $w_{ci}^{-1} = (1 - p_i) q_i$, and $w_{di}^{-1} = (1 - p_i)(1 - q_i)$; with p_i the probability that genotype 1 has a band at position *i*; and q_i the

TABLE 1

Test statistics for scoring range 50–500 bp and an AFLP procedure with *A/C* selective nucleotides

No. bands	No. scored	Exp. bands	95%	99%	Exp. Dice	95%	99%	Max.	Exp. Nei and Li	95%	99%	Max.
10	10	0.39 ± 0.60	1.06	1.89	0.039 ± 0.060	0.106	0.189	1.000	0.961 ± 0.060	0.894	0.811	0.000
10	20	0.78 ± 0.83	1.89	2.81	0.052 ± 0.056	0.126	0.187	0.667	0.948 ± 0.056	0.874	0.813	0.333
10	30	1.16 ± 1.00	2.61	3.56	0.058 ± 0.050	0.131	0.178	0.500	0.942 ± 0.050	0.869	0.822	0.500
10	40	1.54 ± 1.12	3.02	3.99	0.062 ± 0.045	0.121	0.160	0.400	0.938 ± 0.045	0.879	0.840	0.600
10	50	1.91 ± 1.22	3.67	4.72	0.064 ± 0.041	0.122	0.157	0.333	0.936 ± 0.041	0.878	0.843	0.667
10	60	2.27 ± 1.30	4.03	5.09	0.065 ± 0.037	0.115	0.145	0.286	0.935 ± 0.037	0.885	0.855	0.714
10	70	2.63 ± 1.37	4.61	5.70	0.066 ± 0.034	0.115	0.142	0.250	0.934 ± 0.034	0.885	0.858	0.750
10	80	2.98 ± 1.43	4.93	5.97	0.066 ± 0.032	0.110	0.133	0.222	0.934 ± 0.032	0.890	0.867	0.778
10	90	3.32 ± 1.47	5.41	6.53	0.066 ± 0.029	0.108	0.131	0.200	0.934 ± 0.029	0.892	0.869	0.800
10	100	3.66 ± 1.50	5.77	6.84	0.067 ± 0.027	0.105	0.124	0.182	0.933 ± 0.027	0.895	0.876	0.818
10	110	3.98 ± 1.52	6.02	7.10	0.066 ± 0.025	0.100	0.118	0.167	0.934 ± 0.025	0.900	0.882	0.833
10	120	4.31 ± 1.55	6.47	7.55	0.066 ± 0.024	0.100	0.116	0.154	0.934 ± 0.024	0.900	0.884	0.846
20	20	1.55 ± 1.15	3.16	4.21	0.078 ± 0.058	0.158	0.211	1.000	0.922 ± 0.058	0.842	0.789	0.000
20	30	2.31 ± 1.38	4.32	5.55	0.092 ± 0.055	0.173	0.222	0.800	0.908 ± 0.055	0.827	0.778	0.200
20	40	3.06 ± 1.55	5.34	6.65	0.102 ± 0.052	0.178	0.222	0.667	0.898 ± 0.052	0.822	0.778	0.333
20	50	3.80 ± 1.69	6.28	7.67	0.108 ± 0.048	0.179	0.219	0.571	0.892 ± 0.048	0.821	0.781	0.429
20	60	4.52 ± 1.81	7.13	8.62	0.113 ± 0.045	0.178	0.215	0.500	0.887 ± 0.045	0.822	0.785	0.500
20	70	5.23 ± 1.90	7.96	9.49	0.116 ± 0.042	0.177	0.211	0.444	0.884 ± 0.042	0.823	0.789	0.556
20	80	5.93 ± 1.98	8.81	10.30	0.119 ± 0.040	0.176	0.206	0.400	0.881 ± 0.040	0.824	0.794	0.600
20	90	6.61 ± 2.04	9.61	10.99	0.120 ± 0.037	0.175	0.200	0.364	0.880 ± 0.037	0.825	0.800	0.636
20	100	7.28 ± 2.09	10.32	11.81	0.121 ± 0.035	0.172	0.197	0.333	0.879 ± 0.035	0.828	0.803	0.667
20	110	7.93 ± 2.12	10.96	12.53	0.122 ± 0.033	0.169	0.193	0.308	0.878 ± 0.033	0.831	0.807	0.692
20	120	8.56 ± 2.15	11.69	13.11	0.122 ± 0.031	0.167	0.187	0.286	0.878 ± 0.031	0.833	0.813	0.714
30	30	3.45 ± 1.65	5.86	7.20	0.115 ± 0.055	0.195	0.240	1.000	0.885 ± 0.055	0.805	0.760	0.000
30	40	4.56 ± 1.86	7.32	8.80	0.130 ± 0.053	0.209	0.251	0.857	0.870 ± 0.053	0.791	0.749	0.143
30	50	5.66 ± 2.03	8.69	10.22	0.141 ± 0.051	0.217	0.256	0.750	0.859 ± 0.051	0.783	0.744	0.250
30	60	6.73 ± 2.17	9.92	11.62	0.150 ± 0.048	0.220	0.258	0.667	0.850 ± 0.048	0.780	0.742	0.333
30	70	7.79 ± 2.28	11.16	12.86	0.156 ± 0.046	0.223	0.257	0.600	0.844 ± 0.046	0.777	0.743	0.400
30	80	8.83 ± 2.37	12.36	14.04	0.161 ± 0.043	0.225	0.255	0.545	0.839 ± 0.043	0.775	0.745	0.455
30	90	9.85 ± 2.45	13.51	15.25	0.164 ± 0.041	0.225	0.254	0.500	0.836 ± 0.041	0.775	0.746	0.500
30	100	10.84 ± 2.51	14.58	16.36	0.167 ± 0.039	0.224	0.252	0.462	0.833 ± 0.039	0.776	0.748	0.538
30	110	11.82 ± 2.55	15.62	17.41	0.169 ± 0.036	0.223	0.249	0.429	0.831 ± 0.036	0.777	0.751	0.571
30	120	12.77 ± 2.59	16.61	18.41	0.170 ± 0.034	0.221	0.245	0.400	0.830 ± 0.034	0.779	0.755	0.600
40	40	6.04 ± 2.10	9.14	10.79	0.151 ± 0.052	0.228	0.270	1.000	0.849 ± 0.052	0.772	0.730	0.000
40	50	7.49 ± 2.29	10.89	12.68	0.166 ± 0.051	0.242	0.282	0.889	0.834 ± 0.051	0.758	0.718	0.111
40	60	8.91 ± 2.45	12.61	14.43	0.178 ± 0.049	0.252	0.289	0.800	0.822 ± 0.049	0.748	0.711	0.200
40	70	10.32 ± 2.58	14.17	16.00	0.188 ± 0.047	0.258	0.291	0.727	0.812 ± 0.047	0.742	0.709	0.273
40	80	11.70 ± 2.69	15.74	17.70	0.195 ± 0.045	0.262	0.295	0.667	0.805 ± 0.045	0.738	0.705	0.333
40	90	13.05 ± 2.77	17.20	19.16	0.201 ± 0.043	0.265	0.295	0.615	0.799 ± 0.043	0.735	0.705	0.385
40	100	14.37 ± 2.84	18.65	20.67	0.205 ± 0.041	0.266	0.295	0.571	0.795 ± 0.041	0.734	0.705	0.429
40	110	15.66 ± 2.90	19.96	21.96	0.209 ± 0.039	0.266	0.293	0.533	0.791 ± 0.039	0.734	0.707	0.467
40	120	16.92 ± 2.93	21.32	23.34	0.211 ± 0.037	0.266	0.292	0.500	0.789 ± 0.037	0.734	0.708	0.500
50	50	9.29 ± 2.50	13.00	14.88	0.186 ± 0.050	0.260	0.298	1.000	0.814 ± 0.050	0.740	0.702	0.000
50	60	11.06 ± 2.67	15.05	17.00	0.201 ± 0.049	0.274	0.309	0.909	0.799 ± 0.049	0.726	0.691	0.091
50	70	12.81 ± 2.82	17.01	19.03	0.214 ± 0.047	0.284	0.317	0.833	0.786 ± 0.047	0.716	0.683	0.167
50	80	14.52 ± 2.94	18.93	21.02	0.223 ± 0.045	0.291	0.323	0.769	0.777 ± 0.045	0.709	0.677	0.231
50	90	16.19 ± 3.04	20.78	22.92	0.231 ± 0.043	0.297	0.327	0.714	0.769 ± 0.043	0.703	0.673	0.286
50	100	17.84 ± 3.12	22.56	24.75	0.238 ± 0.042	0.301	0.330	0.667	0.762 ± 0.042	0.699	0.670	0.333
50	110	19.44 ± 3.18	24.25	26.48	0.243 ± 0.040	0.303	0.331	0.625	0.757 ± 0.040	0.697	0.669	0.375
50	120	21.02 ± 3.23	25.87	28.06	0.247 ± 0.038	0.304	0.330	0.588	0.753 ± 0.038	0.696	0.670	0.412
60	60	13.18 ± 2.87	17.54	19.63	0.220 ± 0.048	0.292	0.327	1.000	0.780 ± 0.048	0.708	0.673	0.000
60	70	15.26 ± 3.02	19.83	21.97	0.235 ± 0.046	0.305	0.338	0.923	0.765 ± 0.046	0.695	0.662	0.077
60	80	17.29 ± 3.15	22.04	24.33	0.247 ± 0.045	0.315	0.348	0.857	0.753 ± 0.045	0.685	0.652	0.143
60	90	19.30 ± 3.26	24.25	26.58	0.257 ± 0.044	0.323	0.354	0.800	0.743 ± 0.044	0.677	0.646	0.200
60	100	21.26 ± 3.35	26.34	28.70	0.266 ± 0.042	0.329	0.359	0.750	0.734 ± 0.042	0.671	0.641	0.250
60	110	23.18 ± 3.41	28.36	30.75	0.273 ± 0.040	0.334	0.362	0.706	0.727 ± 0.040	0.666	0.638	0.294
60	120	25.05 ± 3.47	30.31	32.71	0.278 ± 0.039	0.337	0.363	0.667	0.722 ± 0.039	0.663	0.637	0.333

(*continued*)

TABLE 1

(Continued)

No. bands	No. scored	Exp. bands	95%	99%	Exp. Dice	95%	99%	Max.	Exp. Nei and Li	95%	99%	Max.
70	70	17.66 ± 3.19	22.52	24.79	0.252 ± 0.046	0.322	0.354	1.000	0.748 ± 0.046	0.678	0.646	0.000
70	80	20.03 ± 3.33	25.06	27.48	0.267 ± 0.044	0.334	0.366	0.933	0.733 ± 0.044	0.666	0.634	0.067
70	90	22.35 ± 3.44	27.61	29.96	0.279 ± 0.043	0.345	0.375	0.875	0.721 ± 0.043	0.655	0.625	0.125
70	100	24.63 ± 3.54	29.98	32.53	0.290 ± 0.042	0.353	0.383	0.824	0.710 ± 0.042	0.647	0.617	0.176
70	110	26.86 ± 3.62	32.37	34.86	0.298 ± 0.040	0.360	0.387	0.778	0.702 ± 0.040	0.640	0.613	0.222
70	120	29.04 ± 3.67	34.65	37.15	0.306 ± 0.039	0.365	0.391	0.737	0.694 ± 0.039	0.635	0.609	0.263
80	80	22.71 ± 3.47	27.97	30.48	0.284 ± 0.043	0.350	0.381	1.000	0.716 ± 0.043	0.650	0.619	0.000
80	90	25.35 ± 3.60	30.83	33.37	0.298 ± 0.042	0.363	0.393	0.941	0.702 ± 0.042	0.637	0.607	0.059
80	100	27.94 ± 3.71	33.61	36.18	0.310 ± 0.041	0.373	0.402	0.889	0.690 ± 0.041	0.627	0.598	0.111
80	110	30.47 ± 3.79	36.26	38.86	0.321 ± 0.040	0.382	0.409	0.842	0.679 ± 0.040	0.618	0.591	0.158
80	120	32.95 ± 3.86	38.83	41.53	0.330 ± 0.039	0.388	0.415	0.800	0.670 ± 0.039	0.612	0.585	0.200
90	90	28.30 ± 3.73	33.97	36.63	0.314 ± 0.041	0.377	0.407	1.000	0.686 ± 0.041	0.623	0.593	0.000
90	100	31.20 ± 3.84	37.04	39.74	0.328 ± 0.040	0.390	0.418	0.947	0.672 ± 0.040	0.610	0.582	0.053
90	110	34.03 ± 3.94	40.01	42.77	0.340 ± 0.039	0.400	0.428	0.900	0.660 ± 0.039	0.600	0.572	0.100
90	120	36.81 ± 4.01	42.92	45.71	0.351 ± 0.038	0.409	0.435	0.857	0.649 ± 0.038	0.591	0.565	0.143
100	100	34.39 ± 3.96	40.47	43.17	0.344 ± 0.040	0.405	0.432	1.000	0.656 ± 0.040	0.595	0.568	0.000
100	110	37.52 ± 4.05	43.74	46.53	0.357 ± 0.039	0.417	0.443	0.952	0.643 ± 0.039	0.583	0.557	0.048
100	120	40.59 ± 4.13	46.91	49.77	0.369 ± 0.038	0.426	0.452	0.909	0.631 ± 0.038	0.574	0.548	0.091
110	110	40.95 ± 4.16	47.33	50.17	0.372 ± 0.038	0.430	0.456	1.000	0.628 ± 0.038	0.570	0.544	0.000
110	120	44.31 ± 4.24	50.81	53.72	0.385 ± 0.037	0.442	0.467	0.957	0.615 ± 0.037	0.558	0.533	0.043
120	120	47.96 ± 4.33	54.61	57.58	0.400 ± 0.036	0.455	0.480	1.000	0.600 ± 0.036	0.545	0.520	0.000

No. bands	No. scored	Exp. Jaccard	95%	99%	Max.	Exp. SM	95%	99%	Min.	Max.
10	10	0.021 ± 0.033	0.056	0.105	1.000	0.957 ± 0.003	0.960	0.964	0.956	1.000
10	20	0.028 ± 0.030	0.068	0.104	0.500	0.937 ± 0.004	0.942	0.946	0.933	0.978
10	30	0.031 ± 0.027	0.070	0.098	0.333	0.916 ± 0.004	0.923	0.927	0.911	0.956
10	40	0.032 ± 0.024	0.064	0.087	0.250	0.896 ± 0.005	0.903	0.907	0.889	0.933
10	50	0.033 ± 0.022	0.065	0.085	0.200	0.875 ± 0.005	0.883	0.888	0.867	0.911
10	60	0.034 ± 0.020	0.061	0.078	0.167	0.855 ± 0.006	0.863	0.867	0.845	0.889
10	70	0.034 ± 0.018	0.061	0.077	0.143	0.834 ± 0.006	0.843	0.848	0.823	0.867
10	80	0.035 ± 0.017	0.058	0.071	0.125	0.814 ± 0.006	0.822	0.827	0.800	0.845
10	90	0.035 ± 0.016	0.057	0.070	0.111	0.793 ± 0.007	0.802	0.807	0.778	0.823
10	100	0.035 ± 0.015	0.055	0.066	0.100	0.772 ± 0.007	0.782	0.786	0.756	0.800
10	110	0.035 ± 0.014	0.053	0.063	0.091	0.752 ± 0.007	0.761	0.765	0.734	0.778
10	120	0.034 ± 0.013	0.052	0.062	0.083	0.731 ± 0.007	0.740	0.745	0.712	0.756
20	20	0.041 ± 0.032	0.086	0.118	1.000	0.918 ± 0.005	0.925	0.930	0.911	1.000
20	30	0.049 ± 0.031	0.095	0.125	0.667	0.899 ± 0.006	0.908	0.914	0.889	0.978
20	40	0.055 ± 0.029	0.098	0.125	0.500	0.881 ± 0.007	0.891	0.896	0.867	0.956
20	50	0.058 ± 0.027	0.099	0.123	0.400	0.862 ± 0.008	0.873	0.879	0.845	0.933
20	60	0.061 ± 0.026	0.098	0.121	0.333	0.843 ± 0.008	0.854	0.861	0.823	0.911
20	70	0.062 ± 0.024	0.097	0.118	0.286	0.824 ± 0.008	0.836	0.843	0.800	0.889
20	80	0.063 ± 0.022	0.097	0.115	0.250	0.805 ± 0.009	0.817	0.824	0.778	0.867
20	90	0.064 ± 0.021	0.096	0.111	0.222	0.785 ± 0.009	0.799	0.805	0.756	0.845
20	100	0.065 ± 0.020	0.094	0.109	0.200	0.766 ± 0.009	0.780	0.786	0.734	0.823
20	110	0.065 ± 0.019	0.092	0.107	0.182	0.747 ± 0.009	0.760	0.767	0.712	0.800
20	120	0.065 ± 0.017	0.091	0.103	0.167	0.728 ± 0.010	0.741	0.748	0.690	0.778
30	30	0.062 ± 0.031	0.108	0.136	1.000	0.882 ± 0.007	0.893	0.899	0.867	1.000
30	40	0.071 ± 0.031	0.117	0.144	0.750	0.865 ± 0.008	0.877	0.884	0.845	0.978
30	50	0.077 ± 0.030	0.122	0.147	0.600	0.848 ± 0.009	0.861	0.868	0.823	0.956
30	60	0.082 ± 0.028	0.124	0.148	0.500	0.830 ± 0.010	0.844	0.852	0.800	0.933
30	70	0.085 ± 0.027	0.126	0.148	0.429	0.813 ± 0.010	0.828	0.835	0.778	0.911
30	80	0.088 ± 0.026	0.127	0.146	0.375	0.795 ± 0.011	0.811	0.818	0.756	0.889
30	90	0.090 ± 0.024	0.127	0.146	0.333	0.778 ± 0.011	0.794	0.802	0.734	0.867
30	100	0.091 ± 0.023	0.126	0.144	0.300	0.760 ± 0.011	0.776	0.784	0.712	0.845
30	110	0.093 ± 0.022	0.126	0.142	0.273	0.742 ± 0.011	0.759	0.767	0.690	0.823

(*continued*)

TABLE 1

(Continued)

No. bands	No. scored	Exp. Jaccard	95%	99%	Max.	Exp. SM	95%	99%	Min.	Max.
30	120	0.093 ± 0.021	0.125	0.140	0.250	0.724 ± 0.011	0.741	0.749	0.667	0.800
40	40	0.082 ± 0.031	0.129	0.156	1.000	0.849 ± 0.009	0.863	0.870	0.823	1.000
40	50	0.092 ± 0.031	0.138	0.164	0.800	0.834 ± 0.010	0.849	0.857	0.800	0.978
40	60	0.099 ± 0.030	0.144	0.169	0.667	0.818 ± 0.011	0.834	0.842	0.778	0.956
40	70	0.104 ± 0.029	0.148	0.170	0.571	0.802 ± 0.011	0.819	0.827	0.756	0.933
40	80	0.109 ± 0.028	0.151	0.173	0.500	0.786 ± 0.012	0.804	0.812	0.734	0.911
40	90	0.112 ± 0.026	0.152	0.173	0.444	0.770 ± 0.012	0.788	0.797	0.712	0.889
40	100	0.115 ± 0.025	0.154	0.173	0.400	0.753 ± 0.013	0.772	0.781	0.690	0.867
40	110	0.117 ± 0.024	0.153	0.171	0.364	0.737 ± 0.013	0.756	0.765	0.667	0.845
40	120	0.119 ± 0.023	0.154	0.171	0.333	0.720 ± 0.013	0.740	0.749	0.645	0.823
50	50	0.103 ± 0.031	0.149	0.175	1.000	0.819 ± 0.011	0.836	0.844	0.778	1.000
50	60	0.113 ± 0.030	0.159	0.183	0.833	0.805 ± 0.012	0.823	0.831	0.756	0.978
50	70	0.120 ± 0.030	0.165	0.189	0.714	0.791 ± 0.013	0.809	0.818	0.734	0.956
50	80	0.126 ± 0.029	0.170	0.193	0.625	0.776 ± 0.013	0.796	0.805	0.712	0.933
50	90	0.131 ± 0.028	0.174	0.196	0.556	0.761 ± 0.013	0.782	0.791	0.690	0.911
50	100	0.136 ± 0.027	0.177	0.198	0.500	0.747 ± 0.014	0.767	0.777	0.667	0.889
50	110	0.139 ± 0.026	0.179	0.198	0.455	0.731 ± 0.014	0.753	0.763	0.645	0.867
50	120	0.142 ± 0.025	0.180	0.198	0.417	0.716 ± 0.014	0.738	0.748	0.623	0.845
60	60	0.124 ± 0.030	0.171	0.196	1.000	0.792 ± 0.013	0.812	0.821	0.734	1.000
60	70	0.134 ± 0.030	0.180	0.203	0.857	0.779 ± 0.013	0.800	0.809	0.712	0.978
60	80	0.142 ± 0.029	0.187	0.210	0.750	0.766 ± 0.014	0.787	0.797	0.690	0.956
60	90	0.148 ± 0.029	0.193	0.215	0.667	0.753 ± 0.014	0.775	0.785	0.667	0.933
60	100	0.154 ± 0.028	0.197	0.219	0.600	0.739 ± 0.015	0.762	0.772	0.645	0.911
60	110	0.158 ± 0.027	0.200	0.221	0.545	0.726 ± 0.015	0.749	0.759	0.623	0.889
60	120	0.162 ± 0.026	0.202	0.222	0.500	0.712 ± 0.015	0.735	0.746	0.601	0.867
70	70	0.145 ± 0.030	0.192	0.215	1.000	0.768 ± 0.014	0.789	0.799	0.690	1.000
70	80	0.155 ± 0.030	0.201	0.224	0.875	0.756 ± 0.015	0.779	0.789	0.667	0.978
70	90	0.163 ± 0.029	0.209	0.230	0.778	0.744 ± 0.015	0.768	0.778	0.645	0.956
70	100	0.170 ± 0.029	0.214	0.237	0.700	0.732 ± 0.016	0.756	0.767	0.623	0.933
70	110	0.176 ± 0.028	0.219	0.240	0.636	0.720 ± 0.016	0.744	0.755	0.601	0.911
70	120	0.181 ± 0.027	0.223	0.243	0.583	0.707 ± 0.016	0.732	0.743	0.579	0.889
80	80	0.166 ± 0.030	0.212	0.235	1.000	0.746 ± 0.015	0.769	0.780	0.645	1.000
80	90	0.176 ± 0.029	0.222	0.244	0.889	0.735 ± 0.016	0.760	0.771	0.623	0.978
80	100	0.184 ± 0.029	0.230	0.252	0.800	0.725 ± 0.016	0.750	0.761	0.601	0.956
80	110	0.192 ± 0.028	0.236	0.257	0.727	0.714 ± 0.017	0.740	0.751	0.579	0.933
80	120	0.198 ± 0.028	0.241	0.262	0.667	0.703 ± 0.017	0.729	0.741	0.557	0.911
90	90	0.187 ± 0.029	0.233	0.255	1.000	0.726 ± 0.017	0.752	0.763	0.601	1.000
90	100	0.197 ± 0.029	0.242	0.265	0.900	0.717 ± 0.017	0.743	0.755	0.579	0.978
90	110	0.206 ± 0.029	0.250	0.272	0.818	0.707 ± 0.017	0.734	0.746	0.557	0.956
90	120	0.213 ± 0.028	0.257	0.278	0.750	0.698 ± 0.018	0.725	0.737	0.534	0.933
100	100	0.208 ± 0.029	0.254	0.275	1.000	0.709 ± 0.018	0.736	0.748	0.557	1.000
100	110	0.218 ± 0.029	0.263	0.285	0.909	0.701 ± 0.018	0.728	0.741	0.534	0.978
100	120	0.227 ± 0.028	0.271	0.292	0.833	0.692 ± 0.018	0.720	0.733	0.512	0.956
110	110	0.229 ± 0.029	0.274	0.295	1.000	0.694 ± 0.018	0.722	0.735	0.512	1.000
110	120	0.239 ± 0.028	0.284	0.305	0.917	0.687 ± 0.019	0.715	0.728	0.490	0.978
120	120	0.250 ± 0.028	0.295	0.316	1.000	0.681 ± 0.019	0.710	0.723	0.468	1.000

Test statistics are based on the Arabidopsis *in silico* AFLP FLD, for AFLP data scored between 50 and 500 bp with *A/C* selective nucleotides. (Top) Columns 1 and 2, band numbers scored in genotypes to be compared (rounded to tens); column 3, expected number of nonidentical bands comigrating across genotypes with standard deviation; columns 4 and 5, 95 and 99% critical values for expected number of nonidentical bands comigrating across genotypes; column 6, expected Dice similarity with standard deviation; columns 7 and 8, 95 and 99% critical values for expected Dice similarity; column 9, maximum possible Dice similarity; columns 10–13, same as columns 6–9, for Nei and Li dissimilarity. (Bottom) Columns 3–6, same as columns 6–9, top, for Jaccard similarity; columns 7–11, same as columns 6–9, top, for simple matching similarity (with addition of minimum possible similarity).

TABLE 2

Numbers of AFLP bands with average numbers of underlying AFLP fragments

	Scoring range >50				Scoring range >100			
Bands	50–400	50–500	50–600	50–700	100–400	100–500	100–600	100–700
10	10.3	10.2	10.2	10.2	10.3	10.2	10.2	10.2
20	21.0	20.9	20.8	20.8	21.0	20.8	20.8	20.7
30	32.2	32.0	31.9	31.8	32.3	31.9	31.8	31.7
40	44.1	43.6	43.4	43.2	44.1	43.5	43.2	43.0
50	56.5	55.7	55.3	55.1	56.6	55.5	55.0	54.7
60	69.6	68.4	67.8	67.5	69.8	68.1	67.4	66.9
70	83.5	81.7	80.9	80.5	83.7	81.4	80.2	79.6
80	98.1	95.7	94.6	93.9	98.5	95.2	93.6	92.8
90	113.6	110.4	108.8	108.0	114.2	109.7	107.6	106.5
100	130.1	125.9	123.8	122.7	131.0	125.0	122.3	120.8
110	147.7	142.1	139.5	138.1	144.0	138.1	135.6	133.8
120	166.4	159.3	156.0	154.2	158.2	153.2	150.7	148.3

Numbers of AFLP bands with average numbers of underlying AFLP fragments, for 12 different numbers of bands and eight scoring ranges. Column 1, number of bands present in an AFLP profile. Columns 2–5, AFLP scoring ranges starting with 50-bp fragments. Columns 6–9, AFLP scoring ranges starting with 100-bp fragments.

probability that genotype 2 has a band at position i. The band probabilities are derived from the fragment probabilities in the Arabidopsis *in silico* AFLP FLD according to $p_i = 1 - [1 - p(\text{fragment at } i)]^{N1}$, and $q_i = 1 - [1 - p(\text{fragment at } i)]^{N2}$, where $N1$ and $N2$ are the total numbers of fragments in the scoring range in genotypes 1 and 2, respectively. The number of fragments N for each genotype depends on the scoring range, the total number of bands within the scoring range, and the fragment length distribution and was determined by Monte Carlo simulation as described in *Significance tests for pairwise AFLP band similarities.* In the second step, weighted similarity coefficients are calculated according to: weighted Dice $= 2a_w/(2a_w + b_w + c_w)$, weighted Jaccard $= a_w/(a_w + b_w + c_w)$, and weighted simple matching $= (a_w + d_w)/(a_w + b_w + c_w + d_w)$. Weighted Nei and Li $= (1 -$ weighted Dice).

The Arabidopsis sequence as a model system: The test statistics in this study are based on *in silico* AFLP FLDs from the Arabidopsis genome sequence. This sequence is generally considered to be representative of the genome of an angiosperm species (*e.g.*, Arabidopsis Genome Initiative 2000; Barnes 2002), and therefore the test statistics based on the Arabidopsis genome sequence should be valid for angiosperms in general.

A limitation of the Arabidopsis sequence is that a significant part is still missing. According to the Arabidopsis Genome Initiative (2000), ~8.5% of the genome has not yet been aligned (~10 of an estimated 125 Mb). This 8.5% mainly consists of repeat sequences in centromeric and rDNA regions. Genetic mapping studies in Arabidopsis (*e.g.*, Alonso-Blanco *et al.* 1998) showed a clustering of AFLP fragments around the centromeres, which could indicate that the actual percentage of AFLP fragments missing from the Arabidopsis AFLP FLD is much higher than the 8.5% of missing sequence. In a recent study, however, Peters *et al.* (2001) found that Arabidopsis *Sac*I/*Mse*I *in silico* AFLP fragments do not cluster around the centromeres, but are evenly dispersed over the genome. They argued that the apparent overrepresentation of AFLP fragments in genetic mapping studies must originate in a higher mutation frequency in the (peri)centromeric regions rather than in an actual overrepresentation of AFLP fragments. Assuming that the findings of Peters *et al.* (2001) are representative for AFLP fragments in general, the missing 8.5% of repeat regions in the Arabidopsis genome sequence corresponds to 8.5% of missing AFLP fragments in the Arabidopsis AFLP FLD. These missing regions contain mainly repeat sequences. Estimating the influence of the missing repeats on the Arabidopsis AFLP FLD is highly speculative, but one could argue that their influence on the significance tests may be only limited. Given the fact that the average size of the individual repeat units is relatively small, the size of AFLP fragments resulting from restriction sites in the repeat regions will also be small. The possible underrepresentation of small fragments will mainly influence the lower part of the Arabidopsis AFLP FLD. In most AFLP studies, these smaller fragments are discarded. Consequently, they do not influence the results.

Specific features of the Arabidopsis genome that may limit its general applicability as a model system for angiosperms are its small size (120 Mb) and its relatively low $G+C$ content (36%). We examined the representativity of the Arabidopsis sequence using sequences of *Oryza sativa* L. Apart from sequences of Arabidopsis, sequences of *O. sativa* L. subspecies *indica* (Yu *et al.* 2002)

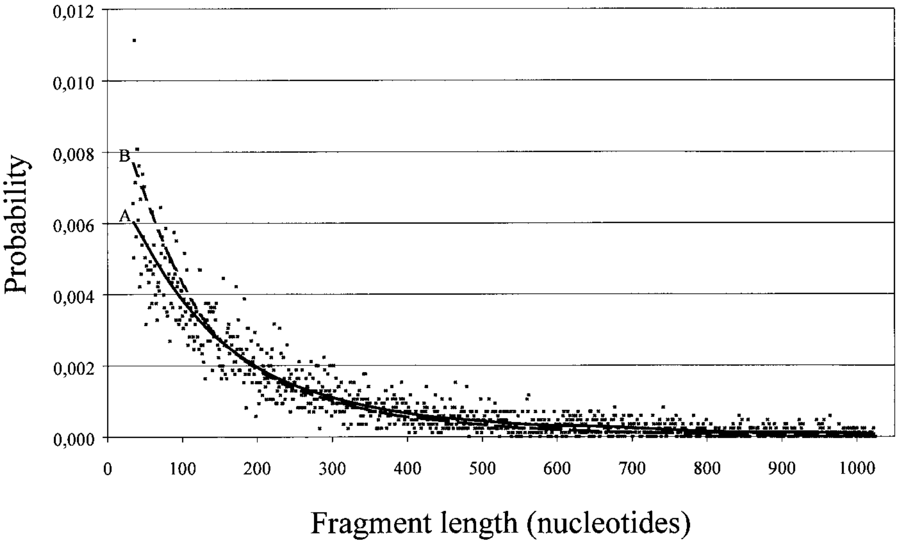

Figure 4.—Relative frequency distribution of fragments resulting from *in silico* AFLP without selective nucleotides on the rice genome sequence (frequencies for each length class are denoted by dots). (A) Smoothed FLD resulting from *in silico* AFLP without selective nucleotides on the rice genome sequence. (B) The smoothed FLD resulting from *in silico* AFLP without selective nucleotides on the Arabidopsis genome sequence is given as a reference. Fragment lengths range from 32 to 1024 bp.

and *japonica* (Feng *et al.* 2002; Goff *et al.* 2002; Sasaki *et al.* 2002) are the only complete angiosperm sequences presently available. However, at the time of our study the *O. sativa* sequences were still very fragmented. We used sequences from chromosomes 3 (43.4% $G + C$) and 10 (43.6% $G + C$) of *O. sativa* subsp. *japonica* (hereafter, rice), covering nearly complete chromosomes contained in a limited number of BAC assemblies. Sequence data were obtained from the web site of The Institute for Genomic Research at http://www.tigr.org. To generate the rice FLD, we performed the *in silico* AFLP as described for Arabidopsis, without selective nucleotides. Vector sequences and sequences of suspect origin were removed from the BAC assemblies prior to *in silico* AFLP, using the National Center for Biotechnology Information VecScreen web tool. The probability distribution of the AFLP fragment lengths was estimated by fitting a cubic smoothing spline as before. The smoothing spline and the relative frequency distribution of the rice *in silico* AFLP fragments are depicted in Figure 4. Fragment sizes range from 32 to 1024 bp.

The Arabidopsis FLD without selective nucleotides is included as a reference. A two-sample Kolmogorov-Smirnov test showed that the rice FLD differs significantly from the Arabidopsis FLDs with A/C, T/A, or without selective nucleotides ($P < 0.0001$), but not from that with C/G selective nucleotides ($P = 0.09$). The most obvious reason for the difference is the high $G + C$ content of the rice sequences relative to those of Arabidopsis. As predicted by the theoretical model of Innan *et al.* (1999), the higher $G + C$ content in rice yields a more even FLD. Additionally, there may be other genome differences between rice and Arabidopsis that influence the AFLP FLD. Most notably, these could be differences related to the evolutionary distinct position of Poaceae within the angiosperms (*e.g.*, Montero *et al.* 1990; Devos *et al.* 1999; Freeling 2001). However, the influence of these additional factors cannot be studied separately from that of $G + C$ content until more evolutionary distinct genome sequences with similar nucleotide compositions become available.

Comparison of the test statistics for Arabidopsis and rice in the scoring range 50–500 bp (supplemental Table 3, available at http://www.dpw.wur.nl/biosys/AFLSIM_UK.html) showed that the expected number of nonidentical bands comigrating across genotypes is on average 10% lower for rice. Although the numbers are in the same order of magnitude, the difference between Arabidopsis and rice illustrates the need for more than one model species. Given the fact that Arabidopsis and rice cover most of the $G + C$ range for angiosperms, together they probably suffice as model species for the angiosperms in general. Therefore, we propose that the tests statistics based on the Arabidopsis sequence be considered generally applicable for angiosperms with $G + C$ contents between ~35 and 40% $G + C$, and tests based on the rice sequence be considered generally applicable for angiosperms with $G + C$ contents between ~40 and 50%. For angiosperms with unknown $G + C$ content, the test statistics for the Arabidopsis genome can be applied as a conservative test. Test statistics based on a more complete rice genome sequence will be developed at a later stage.

DISCUSSION

Theoretical and *in silico* AFLP FLDs were examined as a basis for significance tests for AFLP similarities. Comparison of the theoretical AFLP FLD of Innan *et al.* (1999) with a FLD based on *in silico* AFLP of the complete Arabidopsis genome sequence demonstrated that the theoretical distribution is not representative of that of an actual genome. This is not in accordance with Vekemans *et al.* (2002), who concluded that the theoretical distribution of Innan *et al.* (1999) was representative of empirical distributions of *Phaseolus lunatus* and *Lolium perenne* in a scoring range between 75 and 450 bp. The difference in conclusions may be explained by (1) errors in the empirical data sets, resulting from the AFLP procedure (discussed previously), and (2) fragment numbers in the empirical data sets (801 and 1599, respectively) being too low to yield a representative FLD. The variation in the FLD resulting from the low numbers of fragments probably obscured systematic differences between the theoretical and empirical distributions. In this study, the Arabidopsis *in silico* AFLP FLDs are based on much larger numbers of fragments (23,556 between 75 and 450 bp), enabling a more detailed comparison. This new comparison demonstrated a clear discrepancy between the theoretical and the *in silico* distributions, indicating that theoretical distributions based on Innan *et al.* (1999) do not adequately describe AFLP FLDs based on an actual genome.

The discrepancy between the theoretical and the *in silico* distribution may be explained by two assumptions made by Innan *et al.* (1999). The first is that of a random nucleotide sequence under the Jukes and Cantor (1969) model. In actual genomes the nucleotides are not randomly distributed, but organized in distinct patterns of dinucleotides and oligonucleotides (Nussinov 1981, 1991). At a larger scale, the genome is organized in isochores, showing large blocks of *G* + *C*-rich sequences alternated by large blocks of more *A* + *T*-rich sequences (Salinas *et al.* 1988; Matassi *et al.* 1989; Montero *et al.* 1990). Moreover, the Jukes and Cantor model assumes equal base frequencies and equal chances on substitution among all nucleotides, while in reality base frequencies are unequal and substitution rates vary. The second assumption that may explain the deviation between the theoretical and the *in silico* distribution is that of nucleotide changes as the sole cause of changes in DNA sequence. Under this second assumption, processes such as insertions and deletions are ignored. Obviously, this is a simplification of the dynamics in actual genomes, as was already noted by Innan *et al.* (1999). Both assumptions introduce restrictions in the model of Innan *et al.* (1999) that may be too limiting to allow for an adequate description of an AFLP FLD.

Our analysis of the Arabidopsis *in silico* AFLP FLD demonstrated that the type of selective nucleotides influences the shape of the distribution. Use of only *G* + *C* nucleotides favors the selection of long fragments over short ones, yielding a relatively even distribution of fragments over length classes. Use of only *A* + *T* nucleotides favors the selection of short fragments over long ones, giving a more asymmetrical distribution. The effect probably results from the isochore structure of the genome in combination with the nucleotide composition of the restriction enzymes. The enzymes employed in this study are a frequent cutter (*Mse*I) and a rare cutter (*Eco*RI). Because *Mse*I cuts are much more frequent than *Eco*RI cuts, the average AFLP fragment size will be determined mainly by the frequency of *Mse*I cuts. The restriction site of *Mse*I contains no *G* + *C* nucleotides, and therefore this enzyme will preferably cut in *A* + *T*-rich isochores. Given the preference of the frequent-cutting *Mse*I enzyme to cut in *A* + *T*-rich isochores, and the fact that the fragment size is inversely proportional to the frequency of cuts, AFLP fragments resulting from *A* + *T*-rich isochores will on average be smaller than fragments resulting from other parts of the genome. Because these fragments originate in *A* + *T*-rich stretches of the genome, the fragments themselves will contain relatively high proportions of *A* + *T* nucleotides. Inversely, fragments resulting from *G* + *C*-rich isochores will on average be longer and contain relatively high proportions of *G* + *C* nucleotides (the relation between fraction *G* + *C* and fragment length in the Arabidopsis *in silico* AFLP data is approximately $G + C = 0.34379 + 0.00012036 \times \text{length}$). Using *T*/*A* selective nucleotides in the AFLP procedure will favor the shorter *A* + *T*-rich sequences over the longer *G* + *C*-rich sequences, yielding an asymmetric AFLP FLD with mainly short sequences. Using *C*/*G* selective nucleotides will favor *G* + *C*-rich sequences, yielding a more even distribution of AFLP fragments over length classes. The FLD resulting from an AFLP procedure with *A*/*C* selective nucleotides did not differ significantly from the FLD generated without selective nucleotides, illustrating that the selective nucleotides effect is avoided when mixed *A* + *T*/*G* + *C* selective nucleotides are used.

On the basis of the Arabidopsis *in silico* AFLP FLDs, the numbers of nonidentical bands comigrating across genotypes were calculated as a basis for significance tests for AFLP similarities. Table 1 shows that the proportion of nonidentical bands comigrating across genotypes increases with the number of bands scored per genotype. When 10 bands are scored in each genotype and *A*/*C* selective nucleotides are used, the proportion of comigrating nonidentical bands is ∼4%. For 30 bands, this proportion is 12%, for 60 bands it is 22%, for 90 bands it is 31%, and for 120 bands it is 40%. The increase results from the fact that the probability for nonidentical AFLP fragments to comigrate at the same position increases with increasing numbers of total fragments. Relative to the proportion of comigrating nonidentical bands for *A*/*C* nucleotides, the proportions for *T*/*A* selective nucleotides are somewhat higher (4, 13, 24, 33, and 42%), while the proportions for *C*/*G* nucleotides are somewhat lower (4, 10, 20, 29, and 37%). However, all are in the same order of magnitude. The differences for the various combinations of selective nucleotides probably result from selection bias due to the isochore structure of the genome and the use of different types of selective nucleotides, as discussed before.

The high numbers of nonidentical comigrating bands apparent from Table 1 and supplemental Table 3 illustrate that overestimation of phenetic or genetic similarities based on AFLP band patterns is a serious problem when 50–100 bands per genotype are scored, as recommended by Vos *et al.* (1995). However, even for lower numbers of bands per genotype, a considerable percentage of comigrating bands are nonidentical. Therefore, overestimation of similarities based on AFLP band patterns cannot be completely ruled out by limiting the number of bands within a scoring range. However, the influence of the overestimation on the final analyses can be diminished by using corrected similarities, or weighted similarities, or by removing from the data sets those genotypes without any significant similarity to other genotypes. This article provides the procedures that enable this, all of which are available in the program AFLSIM. The procedures can be applied in, *e.g.*, genetic diversity studies or phylogenetic studies, which often include less-related genotypes as reference groups. For any genotype to be useful as a reference, at least some genetic similarity with the group under study is required. In many genetic diversity studies, however, the genetic similarities between the groups under study and the reference group are below the 95% critical values indicated in our tests. Such similarities, usually in the order of 0.15 or 0.20, are mistakenly taken to indicate a proper level of similarity for a reference group. To select a proper reference group, pairwise similarities between genotypes in the reference group and in the group under study should be tested, and at least some similarities between genotypes of both groups should be significant. Reference genotypes without significant similarity to the group under study should be discarded prior to further analysis.

By enabling the detection of unrelated genotypes and by the use of corrected and weighted similarity values, application of the procedures proposed in this article will make the analysis of AFLP data sets more informative and more reliable.

We thank Ronald van den Berg of the Biosystematics Group, Wageningen University, The Netherlands; Hans de Jong of the Laboratory of Genetics, Wageningen University, The Netherlands; and two anonymous reviewers for useful suggestions. Herman van Eck of the laboratory of Plant Breeding, Wageningen University, The Netherlands, is acknowledged for discussing the theoretical aspects of the AFLP technique; and Qiaoping Yuan of The Institute of Genome Research (TIGR) is acknowledged for providing additional information on the Arabidopsis and rice genome sequences. Sequence data of Arabidopsis and rice were obtained from TIGR through their web site at http://www.tigr.org.

LITERATURE CITED

Alonso-Blanco, C., A. J. M. Peeters, M. Koornneef, C. Lister, C. Dean *et al.*, 1998 Development of an AFLP based linkage map of L*er*, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a L*er*/Cvi recombinant inbred line population. Plant J. **14:** 259–271.

Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815.

Barnes, S., 2002 Comparing Arabidopsis to other flowering plants. Curr. Opin. Plant Biol. **4:** 1–6.

Barow, M., and A. Meister, 2002 Lack of correlation between AT frequency and genome size in higher plants and the effect of nonrandomness of base sequences on dye binding. Cytometry **47:** 1–7.

Devos, K. M., J. Beales, Y. Nagamura and T. Sasaki, 1999 Arabidopsis-rice: Will colinearity allow gene prediction across the eudicot-monocot divide? Genome Res. **9:** 825–829.

Dice, L. R., 1945 Measures of the amount of ecological association between species. Ecology **26:** 297–302.

El-Rabey, H. A., A. Badr, R. Schafer-Pregl, W. Martin and F. Salamini, 2002 Speciation and species separation in *Hordeum* L. (Poaceae) resolved by discontinuous molecular markers. Plant Biol. **4:** 567–575.

Feng, Q., Y. Zhang, P. Hao, S. Wang, G. Fu *et al.*, 2002 Sequence and analysis of rice chromosome 4. Nature **420:** 316–320.

Freeling, M., 2001 Grasses as a single genetic system. Reassessment 2001. Plant Physiol. **125:** 1191–1197.

Goff, S. A., D. Ricke, T. H. Lan, G. Presting, R. Wang *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science **296:** 92–100.

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

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

Jaccard, P., 1908 Nouvelles recherches sur la distribution florale. Bull. Soc. Vaudoise Sci. Nat. **44:** 223–270.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Karp, A., O. Seberg and M. Buiatti, 1996 Molecular techniques in the assessment of botanical diversity. Ann. Bot. **78:** 143–149.

Marie, D., and S. C. Brown, 1993 A cytometric exercise in plant DNA histograms, with 2C values for 70 species. Biol. Cell **78:** 41–51.

Matassi, G., L. M. Montero, J. Salinas and G. Bernardi, 1989 The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. Nucleic Acids Res. **17:** 5273–5290.

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

Montero, L. M., J. Salinas, G. Matassi and G. Bernardi, 1990 Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res. **18:** 1859–1867.

Mueller, U. G., and L. LaReesa Wolfenbarger, 1999 AFLP genotyping and fingerprinting. Trends Ecol. Evol. **14:** 389–394.

Munford, A. G., 1977 A note on the uniformity assumption in the birthday problem. Am. Stat. **31:** 119.

Nagl, W., and B. Stein, 1989 DNA characterization in host-specific *Viscum album* subspecies (Viscaceae). Plant Syst. Evol. **166:** 243–248.

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

Nussinov, R., 1981 Nearest neighbor nucleotide patterns. J. Biol. Chem. **256:** 8458–8462.

Nussinov, R., 1991 Compositional variations in DNA sequences. Comput. Appl. Biosci. **7:** 287–293.

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

Peters, J. L., H. Constandt, P. Neyt, G. Cnops, J. Zethof *et al.*, 2001 A physical amplified fragment-length polymorphism map of *Arabidopsis*. Plant Physiol. **127:** 1579–1589.

Rohlf, F. J., 1993 *NTSYS-pc, Numerical Taxonomy and Multivariate Analysis System.* Exeter Software, Setauket, NY.

Rouppe van der Voort, J. N. A. M., P. Van Zandvoort, H. J. Van Eck, R. T. Folkertsma, R. C. B. Hutten *et al.*, 1997 Use of allele specificity of comigrating AFLP markers to align genetic maps from different potato genotypes. Mol. Gen. Genet. **255:** 438–447.

Salinas, J., G. Matassi, L. M. Montero and G. Bernardi, 1988 Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucleic Acids Res. **16:** 4269–4285.

Sasaki, T., T. Matsumoto, K. Yamamoto, K. Sakata, T. Baba *et al.*, 2002 The genome sequence and structure of rice chromosome 1. Nature **420:** 312–316.

Sokal, R. R., and P. H. A. Sneath, 1963 *Principles of Numerical Taxonomy.* W. H. Freeman, San Francisco/London.

Vekemans, X., T. Beauwens, M. Lemaire and I. Roldan-Ruiz, 2002 Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Mol. Ecol. **11:** 139–151.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. Van de Lee *et al.*, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23:** 4407–4414.

Yu, J., S. N. Hu, J. Wang, K. S. G. Wong, S. G. Li *et al.*, 2002 A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:** 79–92.

Communicating editor: M. A. F. Noor