On some surprising statistical properties of a DNA fingerprinting technique called AFLP

Gerrit Gort

Thesis committee

Thesis supervisors
Prof. dr. ir. A. Stein
Professor of Mathematical and Statistical Methods
Biometris, Wageningen University
Presently
Professor of Mathematical and Statistical Methods for Geodata
ITC, Enschede

Prof. dr. F.A. van Eeuwijk
Professor of Applied Statistics
Biometris, Wageningen University

Other members
Prof. dr. ir. J.A.M. van Arendonk, Wageningen University
Prof. dr. R.C. Jansen, University of Groningen
Dr. ir. C.A. Maliepaard, Wageningen University
Prof. dr. H.P. Piepho, University of Hohenheim, Germany

This research was conducted under the auspices of the "C.T. de Wit Graduate School for Production Ecology & Resource Conservation".

On some surprising statistical properties of a DNA fingerprinting technique called AFLP

Gerrit Gort

Thesis
submitted in partial fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Doctorate Board
to be defended in public
on Wednesday 20 January 2010
at 4 PM in the Aula.

Gerrit Gort
On some surprising statistical properties of a DNA fingerprinting technique called AFLP, 164 pages.

Thesis Wageningen University, Wageningen, NL (2010)
With references, with summaries in Dutch and English

ISBN 978-90-8585-537-8

Abstract

AFLP is a widely used DNA fingerprinting technique, resulting in band absence - presence profiles, like a bar code. Bands represent DNA fragments, sampled from the genome of an individual plant or other organism. The DNA fragments travel through a lane of an electrophoretic gel or microcapillary system, and are separated by length, with shorter fragments traveling further. Multiple individuals are simultaneously fingerprinted on a gel. One of the applications of AFLP is the estimation of genetic similarity between individuals, e.g. in diversity and phylogenetic studies. In that case, profiles of two individuals are compared, and the fraction of shared (comigrating) bands is calculated, e.g. using the Dice similarity coefficient. Two comigrating bands may share the same fragment, but band sharing could also be due to chance, if two equally sized, but different fragments are amplified. This is called homoplasy. Homoplasy biases similarity coefficients. Homoplasy could also occur within a lane, if two different fragments of equal length are amplified, resulting in a single band. We call this collision. The main objective in this thesis is the study of collision and homoplasy in AFLP. The length distribution of AFLP fragments plays an important role. This distribution is highly skewed with more abundant short fragments. By simulation the expected similarity for unrelated genotypes is calculated. As much as 40% of the bands may be shared by chance in case of profiles with 120 bands. The collision problem is analogous to the birthday problem, which has a surprising solution. The collision problem is even more extreme, making it even more surprising. Profiles with only 19 bands contain collision(s) with probability $> 1/2$. These findings have consequences for practice. In some cases it is better to prevent the occurrence of collisions by decreasing the number of bands, in other cases a correction for homoplasy and collision is preferred. Modified similarity coefficients are proposed, that estimate the fraction of homologous fragments, correcting for homoplasy and collision. Partially related to homoplasy and collision, we study the codominant scoring of AFLP in association panels. Examples of AFLP in lettuce and tomato serve as illustrations.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

1.1 Introduction

AFLP® is a DNA fingerprinting technique, developed and patented by Keygene N.V. [1]. The seminal paper on AFLP is Vos et al. (1995), which has been cited as of date over 5000 times [2]. Although not explicitly stated in this paper, the name is interpreted as an acronym of Amplified Fragment Length Polymorphism, giving an indication of the working of AFLP: it aims to find differences (polymorphisms) in lengths of DNA fragments, which were copied (amplified) many times.

AFLP is used in many fields of the life sciences. We studied AFLP in cooperation with plant scientists working in Taxonomy, Genetic Resources, and Breeding. In the taxonomic study the aim was to infer species relationships in lettuce; in the Genetic Resources study AFLP was used as a tool for assessment of diversity and for genebank management; the Breeding study was an association study in tomato, relating phenotypic to genotypic data. And indeed, the majority of applications of AFLP are found in the plant sciences (e.g. Koopman, Zevenbergen, & Van den Berg, 2001), but AFLP is also regularly used in the animal sciences (e.g. Dasmahapatra, Hoffman, & Amos, 2009), in microbiology (e.g. Duim et al., 2001), and to a lesser extent in human genetics (e.g. Prochazka, Walder, & Xia, 2001). AFLP has become a popular tool for genetic relationship, diversity and population genetic studies in many settings (e.g. nature conservation and gene banks), but is used in many other studies as well, ranging from variety identification and marker-assisted breeding, to genetic map construction and QTL-studies, and criminal and paternity tests. A variant of AFLP is cDNA-AFLP (e.g. Breyne et al., 2003), rendering insight into gene expression.

Originating from the early nineties of the 20th century, in the dynamic era of genetics and bioinformatics, AFLP, at the age of 16 since its year of patent, may be considered quite old. The title of the review paper by Meudt and Clarke (2007) "Almost Forgotten or Latest Practice?" suggests the same. A simple way to check the present scope of AFLP is to count the number of publications, making

[1] We will omit the ® sign from here

[2] Source: Web of Science, 15-09-2009

mention of it. Figure 1.1 shows the yearly number of scientific papers referring to the AFLP procedure. The figure demonstrates that the application of AFLP, after a quick rise around the change of the century, currently remains at a constant, high level. Notice that in recent years researchers tend to mention AFLP in the title of their publications less than before, although the total number of papers remains constant.

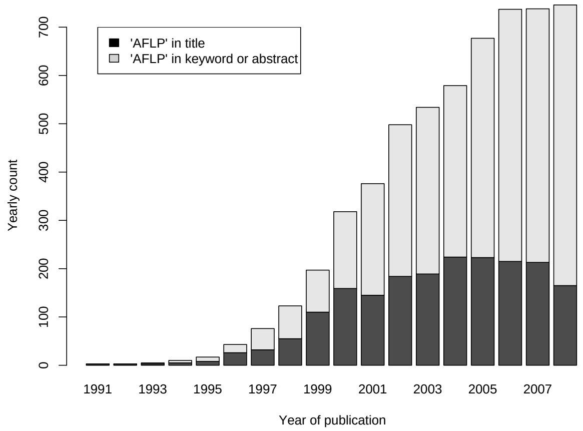


Figure 1.1: Yearly counts of scientific publications in Web of Science in period 1991-2008, containing the phrase 'AFLP' or 'Amplified Fragment Length Polymorphism' in title, keyword or abstract.

Having sketched the place and scope of AFLP in the field of the life sciences, we introduce the topic of this thesis. The result of AFLP is a pattern of bands, like bar codes, in different lanes of an electrophoretic gel or microcapillary system (for details see section 1.2). A band is supposed to represent a DNA fragment. Corresponding bands in different lanes are supposed to be homologous, that is, the DNA fragments are identical and originate from the same genomic locus. But the problem is, that what you see, may not be what you get. Within a lane you see one band, but you may get more than one fragment. Comparing two lanes, you see two identical bands, but you may get two different fragments. The main topic of this thesis is the study of these two problems from a probabilistic and statistical point of view. In sections 1.3 and 1.4 the problems are introduced in greater detail.

1.2 AFLP: technique and data

To understand the ideas in this thesis, some insight into the AFLP technique at a biomolecular level is useful. The AFLP technique consists of four steps: 1) restriction of DNA, 2) ligation of adaptors, 3) amplification of fragments, and 4) visualization of amplified fragments. Below we explain these steps, and give them, where relevant, a statistical interpretation.

1. For each individual the total DNA is cut into fragments using two molecular "scissors" (restriction enzymes). The restriction enzymes recognize specific nucleotide sequences within the DNA strands, and cleave them at these sites. Usually a short-cutter, with short restriction site (most commonly *Mse*I with restriction site $TTAA$), and a long-cutter (most commonly *Eco*RI with restriction site $GAATTC$) are used. The result is a huge collection of DNA fragments of different lengths, typically between 50 and 600 nucleotides. We notice that, from a statistical point of view, for each genome a population of fragments is created, which we call the *population of candidate fragments.* The frequency distribution of the lengths of all fragments in this population is called the *fragment length distribution* (fld).
2. Before a selection of fragments can be taken, some preparatory work is done: two double-stranded DNA sequences (adaptors), that recognize the restriction sites, are glued (ligated) to the ends of the fragments. These adaptors will be used as docking stations for the primers in the next step.
3. A selection of fragments is taken. Only the fragments in this selection will be copied many times (amplified) by the Polymerase Chain Reaction (PCR), so that they will be visible in the next step. Both selection and amplification are caused by primers, which are single-stranded DNA sequences, specific for the two adaptors. The primers bind to the sequence of adaptors, restriction sites and small number of nucleotides adjacent to the restriction sites. Because the primers have selective nucleotides, only fragments with nucleotides complementary to these selective nucleotides are selected for amplification by PCR. Each extra selective nucleotide will cause a reduction in number of fragments of approximately factor 4. The set of two primers is called a primer combination (pc). Only the primer corresponding to the long-cutter restriction enzyme is labeled with a fluorescent dye or by radioactivity. Hence only fragments with at least one *Eco*RI restriction site are eligible for visualization. From a statistical point of view, a *sample* of fragments is drawn from the population of candidate fragments. The sample size is determined by the number of selective nucleotides of the primers, but also by the genome size itself. A larger genome will in general result in a larger sample. Researchers generally strive for a sample size of 50-100 bands per lane (see next step).
4. Separation of the amplified fragments on a gel or microcapillary system. Here we describe only the fragment separation on a gel. On a typical gel, amplified fragments from up to 48 genotypes are separated simultaneously in different columns, or lanes. The fragments are separated by electrophoresis. The main driving force for separation is the *length* of a fragment, with smaller fragments traveling further within a lane. The labeled fragments become visible as bands. Inclusion of a ladder of DNA fragments with known lengths in one of the lanes enables determination of the length of the fragments. Typically only fragments with lengths between 50 and 600 base pairs (bp) are scored. This is called the scoring range. Statistically speaking, the fragment length distribution is truncated: only lengths within the scoring range are observed, smaller and larger fragments are discarded.

The result of these 4 steps is a gel showing bands in different positions within lanes, where the lanes correspond to the different genomes. The pattern of bands within a lane is called a profile. An example of such an AFLP gel is given in figure 1.2, which shows profiles of 47 tomato cultivars. Sometimes the word "AFLP" is used as an alternative for the word "profile". A band from a profile is also called an AFLP marker, or just an AFLP. The word AFLP can therefore have different interpretations: it could mean the technique itself, or a profile from the gel, or a band from the profile. In general the meaning should be clear from the context, and we will not try to be overly precise here.

Next, the band information on the gel is scored. Band information is usually scored binary, that is, bands at a specific position are either absent or present, without interpreting the intensity of the bands. In this way, AFLP markers are dominant, anonymous markers. Dominant scoring of AFLPs means that each fragment is scored as either present or absent, so that a heterozygous genotype cannot be discriminated from a homozygous genotype. Scoring as anonymous markers means that the fragments are recognized only by their length, while their nucleotide sequence remains largely unknown. Statistically speaking, the scoring after step 4 leads to binary information on bands: bands corresponding to fragments with lengths between, say 50 and 600, are either absent or present.

For more detailed information about the AFLP technique we refer to Mueller and LaReesa Wolfenbarger (1999), and Blears, De Grandis, Lee, and Trevors (1998). More recent papers on the AFLP technique are the protocol description by Vuylsteke (2007), and the review by Meudt and Clarke (2007).

1.3 Pros and problems in AFLP

AFLP is said to have a number of advantageous characteristics compared to other DNA fingerprinting techniques:

1. AFLP is highly sensitive and reproducible;
2. No prior sequence information is needed for amplification, making the technique very useful in the study of taxa with little knowledge about the genomic makeup;
3. AFLP has the capability to detect various polymorphisms in different genomic regions simultaneously: up to 100 bands or more per lane may be scored.

Because of these properties, AFLP has become an established DNA fingerprinting technique. This, obviously, does not mean that AFLPs are without flaws. During the process from DNA extraction to interpretation of AFLP profiles, many things may go wrong. Here follows a compilation of possible problems, without trying to be complete:

1. During the preparation of the DNA for AFLP fingerprinting contamination with strange DNA may occur; in that case not all bands on the gel represent the genome under study;
2. During the generation of AFLP profiles, technical problems may occur; for example, variation in fragment mobility may result in bands which do not comigrate correctly, variation in fragment amplification may cause bands to be too vague to be scored properly;

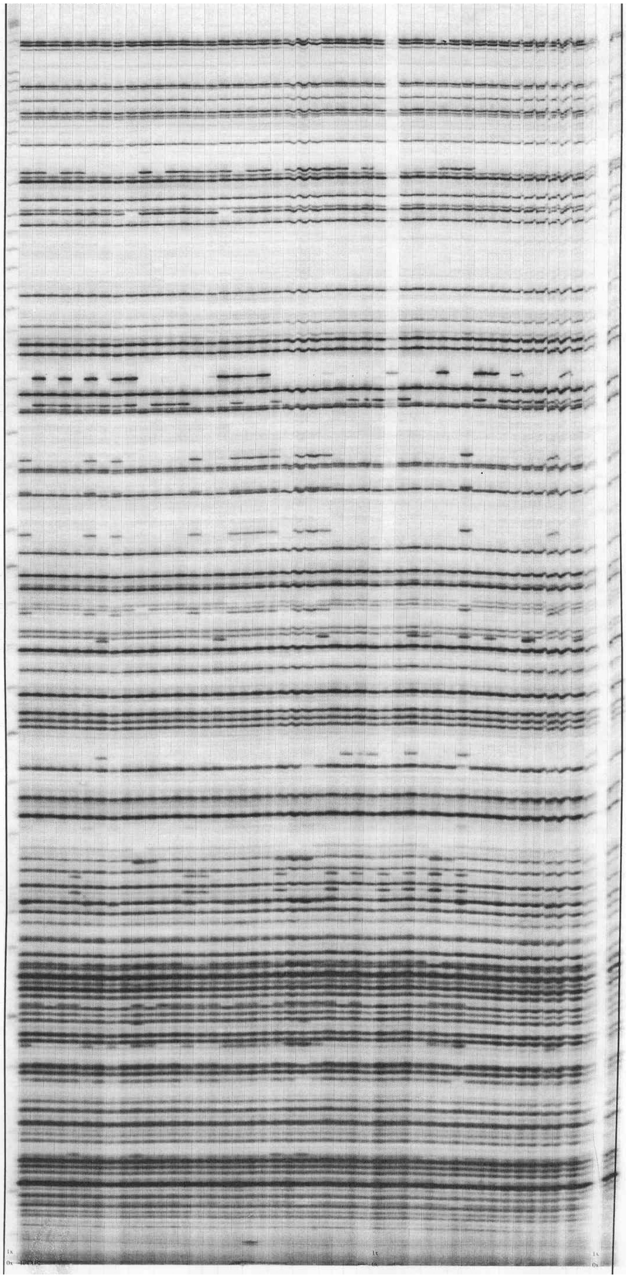


Figure 1.2: Example of an AFLP gel, showing lanes corresponding to 47 tomato cultivars, genotyped within the Center of Biosystems Genomics; restriction enzymes *Eco*RI and *Mse*I were used, and primers with 3 selective nucleotides; the first lane contains a ladder of DNA fragments of known length; fragments at the bottom of the gel have lengths close to 50 bp, at the top 500 bp.

3. During the generation of the binary (0,1) matrix from raw AFLP profiles, subjectivity and human error may play a role, if scoring is not done automatically. And even if scoring is automated, uncertainty remains. Which intensity threshold should be used for bands to be scored as present? Which bin-width should be used for bands to be considered homologous? What is the minimum fragment size to be scored?
4. The interpretation of the binary (0,1) matrix of bands as homologous DNA fragments is troublesome. Equally sized fragments from different genomic loci may have comigrated in two lanes, and the resulting bands may erroneously be interpreted as homologous fragments. This problem is called band size homoplasy, or simply *homoplasy*. Comigration of equally sized, but different fragments may also occur *within* a single lane. We call this problem *collision*, because two or more fragments "collide" in a single band. The problems of homoplasy and collision form the core of the present study.

1.4 Collision and homoplasy

The interpretation of the binary band information as absent / present information of single DNA fragments is problematic. Within a lane comigration of different fragments of equal size may have occurred, leading to multiple fragments within a single AFLP band (collision), which nevertheless is usually interpreted as a single fragment. Comigration of different fragments of equal size in two or more lanes may have occurred, leading to the problem of homoplasy: the erroneous interpretation of bands as homologous.

In the AFLP literature, the problem of homoplasy is well recognized. We cite a few authors here: "Homoplasy is a major issue in the analysis and interpretation of AFLP data." (Meudt & Clarke, 2007), and "Two types of error prevail in AFLP genotyping: allele homoplasy and scoring errors." (Bonin, Ehrich, & Manel, 2007). Meudt and Clarke (2007) also mention: "The quantification of homoplasy in many AFLP datasets both experimentally and via simulation, as well as identification of potential effects that homoplasy might have on results, are key research directions that require further study.".

The problem of collision is less well recognized in the literature. It is, if at all, treated under the heading of homoplasy, and called size homoplasy (Vekemans, Beauwens, Lemaire, & Roldán-Ruiz, 2002) or described as masking (Meudt & Clarke, 2007).

In the literature various ways of assessing homoplasy or collisions are described:

1. In-silico AFLP and Monte Carlo simulation. In-silico AFLP can be performed for species with sequenced genomes, mimicking the AFLP procedure on the computer. In Monte Carlo simulation studies, AFLP is simulated by sampling from a given fragment length distribution. No physical AFLP profiles are created. Studies of this type include
   a) Vekemans et al. (2002)
   b) Althoff, Gitzendanner, and Segrave (2007).
2. Single nucleotide primer extension. The starting point is an AFLP profile resulting from a given primer combination. Next, four extra profiles are made, using the same primer combination, but with one primer extended with a single

nucleotide A, C, T, or G. The four resulting profiles are compared with the original profile. Bands found in more than one of the four extra profiles suggest collision or homoplasy. Realize that fragments with an equal extra nucleotide, not necessarily are identical: there may still be differences in the second next nucleotide or further. Therefore, this method must give a lower bound of problematic bands. Studies of this type include
   a) Hansen, Kraft, Christiansson, and Nilsson (1999)
   b) O'Hanlon and Peakall (2000a)

3. Sequencing of fragments. In studies of this type, AFLP bands are cut out of the gel, re-amplified and cloned into bacteria, which form colonies. A number of colonies are selected, and the bacterial plasmid DNA is sequenced, resulting in the nucleotide sequence of the captured AFLP fragments. The numbers of clones per band differ between studies. Sometimes only a few clones (as low as two) are taken, making it doubtful whether all fragments are sequenced in case of collision. Usually the studies report the extent of sequence identity between fragments, expressed as percentages. Studies of this type include
   a) Rouppe van der Voort et al. (1997)
   b) Meksem, Ruben, Hyten, Triwitayakorn, and Lightfoot (2001)
   c) El-Rabey, Badr, Schafer-Pregl, Martin, and Salamini (2002)
   d) Mechanda, Baum, Johnson, and Arnason (2004)
   e) Mendelson and Shaw (2005)
   f) Ipek, Ipek, and Simon (2006)

To get an idea of the extent of the problems of homoplasy and collision, we summarize some results of the studies mentioned above.

1. Monte Carlo and in-silico AFLP studies
   a) Vekemans et al. (2002) describe the AFLP fingerprinting of lima bean (*Phaseolus lunatus*) and perennial ryegrass (*Lolium perenne*). Using Monte Carlo simulation, they conclude that 250 fragments are needed to get 167.0 bands, close to the observed ±169.3 bands (average total number per primer pair) over all 50 individuals of the first species, and 220 fragments to get 160.0 bands(close to the observed $\pm 154 - 163$ bands over all $30 - 31$ individuals of the second species. This means that 33% and 27% of the fragments are masked. Per individual, on average 150 fragments are needed to get close to the average ±115.3 bands per primer pair in *P. Lunatus* (23% masked), and 80 fragments to get close to the average ±70.2 bands in *L. perenne* (12% masked).
   b) Althoff et al. (2007) use in-silico AFLP on sequenced genomes from 8 wildly diverse organisms (from *Bacillus anthracis* to *Homo sapiens*). For small genomes (with up to 34 bands per profile) the average percentage of bands without collisions is 89%, decreasing to 50% for large genomes (with up to an unrealistic 182 bands per profile). Homology of bands for closely related organisms is very high (100%), but for less related organisms it can have any value between 0% and 100%.
2. Single nucleotide primer extension
   a) Hansen et al. (1999) evaluates AFLP in beet (*Beta*). Part of the study involves AFLPs with extra selective nucleotides for 8 pc's in 2 genotypes. Of the 456 investigated bands, 60 (13.2%) contain at least two fragments

(collision).

b) O'Hanlon and Peakall (2000a) evaluates the method of extra selective nucleotides in *Carduinae* thistles. Out of 94 bands 3 bands are amplified by more than one extra primer (collision). However, it is not clear from how many profiles the 94 bands originate. Of 91 fragments shared between samples, 53% has a different nucleotide at the analyzed position. For closely related individuals the average size homoplasy was only 2.5%, whereas for more distantly related individuals it was as high as 100%.

3. Sequencing of fragments

   a) Rouppe van der Voort et al. (1997) study comigrating AFLP markers for map alignment in potato (*Solanum tuberosum*). In total 733 segregating bands from 12 pc's in 5 parental lines are selected, and 131 comigrating AFLP markers are identified, from which 117 map to the same genomic region. In this group 20 are selected for sequencing, resulting in 5 markers with identical sequences, 13 having up to 10 different nucleotides (of which 2 pairs are allelic, differing 1 bp in size), 1 with a variable stretch of 46 nucleotides, and 1 reported as not homologous. In passing, it is mentioned that occasionally collision was observed. We conclude that even for selected comigrating markers, mapping to the same position, homoplasy is found.

   b) Meksem et al. (2001) aim at the conversion of AFLP bands into high-throughput DNA markers in soybean (*Glycine max*). Six AFLP bands are selected, and 4-30 clones per band sequenced. An astonishing 6 sequences per band on average are found, with approximately the same size.

   c) El-Rabey et al. (2002) perform a phylogenetic study in barley (*Hordeum*) with 63 accessions of 9 species. Five pc's are used, resulting in 906 polymorphic bands. In two groups of comigrating bands across species the sequence identity is determined: group 1 (perfectly aligned bands with same intensity) shows 100% identity within species, and 82.1-100% between species, and group 2 (closely aligned band with different intensities) shows often $< 40\%$ identity between species. No mention of collision is made. We conclude that sequence identity depends on phylogenetic distance, but also on physical characteristics of the bands.

   d) Mechanda et al. (2004) perform a sequencing study of (only) 2 AFLP bands at 4 taxonomic levels (genus, species, variety, population) of *Echinacea*: 1 monomorphic band (273 bp, 79 individuals) and 1 polymorphic band (159 bp, 48 individuals). For the monomorphic band the sequence identity within population is $> 90\%$, within variety $83-95\%$, within species $76-99\%$, within genus 59%. For the polymorphic band the sequence identity is considerably less. Even two clones from the same band may be different in size and sequence (collision). Identity within sample is $52-100\%$, within variety $33-100\%$, within species $24-45\%$, and within genus 1.25%. The conclusion from the authors is that in general comigrating bands cannot be considered homologous.

   e) Mendelson and Shaw (2005), as part of a review paper about AFLP in arthropods, describe a pilot study about the homology of 8 sets of same-sized bands in crickets (*Laupala*). Seven out of 8 bands are confirmed to be homologous.

f) Ipek et al. (2006) study sequence homology of 7 polymorphic AFLP markers in 37 garlic varieties (*Allium sativum* L.), using two pc's. In total 87 bands are sequenced from 4-27 varieties. Per band 2-4 bacterial colonies for sequencing are taken, resulting in 191 amplicons, to yield 124 different fragments. For all 7 markers in at least one of the varieties collisions are found, ranging from only 1 variety out of 12, to 7 out of 17. Up to 4 fragments are found within a single band. For 4 out of 7 markers all varieties share a single fragment. For the remaining 3 markers different levels of homoplasy are found. Up to 17 different fragments (collisions included) are traced. Not all different sequences from one band have exactly equal lengths.

The general conclusions we draw from these studies are: 1) collision occurs regularly, although not always reported; this may be partly due to insufficient sequencing efforts; 2) homoplasy occurs regularly with larger rates for more distantly related individuals; 3) homologous fragments not necessarily have 100% sequence identity.

All the described studies are case studies on specific organisms, or sets of organisms, and lack generality. Some of them study collision, others focus on homoplasy, still others touch upon both. Our work adds to the already extensive literature by modeling AFLP from a statistical point of view. Using a modeling approach, we are able to *estimate* the level of collision or homoplasy. Generalization brings as benefits the allowance of prediction of collision and homoplasy in AFLP in any other case, and formulation of corrected version of derived quantities, like corrected similarity coefficients.

1.5 Objectives

1. The main objective is the study of collision and homoplasy in AFLP, focusing on quantification of the problem, consequences for practice, and correction for derived quantities. To this end we formulate a number of partial objectives:
   a) Estimate fragment length distributions;
   b) Derive critical values of numbers of shared bands and similarity coefficients in case of unrelated genotypes for hypothesis testing;
   c) Derive the probability of at least one collision in a lane in three situations: given the fragment count, the band count, and the band positions;
   d) Estimate the collision count for a lane in three situations: given the fragment count, the band count, and the band positions;
   e) Derive the collision probability for an individual band in three situations: given the fragment count, the band count, and the band positions;
   f) Derive modified similarity coefficients, corrected for homoplasy and collision, and their properties;
2. A second objective is the study of codominant scoring in association panels.

1.6 Outline of the thesis

In this thesis we present a number of papers which we have written over the past few years on the problems of collision and homoplasy, and on codominant scoring in AFLP. Each chapter can be read as an independent item. Therefore, some overlap exists between the chapters, e.g. each chapter starts with a short description of AFLP.

Chapter 2: Significance Tests and Weighted Values for AFLP Similarities, Based on Arabidopsis in Silico AFLP Fragment Length Distributions
This paper, published in Genetics, arose as result of joint work with Wim Koopman, who as a PhD student was using AFLP fingerprints to study the taxonomy of *Lactuca* species. He wanted to know which values of similarity coefficients like Dice or Jaccard, calculated from binary AFLP data, would indicate phylogenetic relationship between genotypes. Thinking about this question, we came across the problems of homoplasy and collision in AFLP. We discovered the connection with probability theory, more specifically the birthday paradox, and the need to have an estimate of the fragment length distribution. We estimated this distribution using an in-silico AFLP approach on the genome sequences of *Arabidopsis thaliana* and *Oryza sativa*, which became available in that time. By simulation we were able to calculate the distribution of similarity coefficients for completely unrelated genotypes. We also wrote a small software program AFLSIM for simulation of AFLP data and calculation of threshold values of similarities. (We erroneously described the Nei-dissimilarity coefficient, as 1-Dice.)

Chapter 3: Fragment length distributions and collision probabilities for AFLP markers
In this paper, published in Biometrics, we took a more formal statistical approach to AFLP. We discussed in more detail how fragment length distributions could be estimated, using a theoretical approach, an in-silico approach, and an empirical data approach. With this last approach the fld is estimated directly from the AFLP profile itself, using a monotonic smoothing spline. For a number of fld's and scoring ranges, we studied the probability distribution of the number of collisions given the number of fragments, or the number of bands in a lane.

Chapter 4: Collision probabilities for AFLP bands, with an application to simple measures of genetic similarity
In the next paper, which was published in the Journal of Agricultural, Biological, and Environmental Statistics, we studied how the collision probability of an individual band depends on the position of the band within the lane (or, equivalently, fragment length), given either the total number of fragments, the total number of bands, or the band positions of all bands in the lane. This is important, because it allows the researcher to assess how trustworthy individual bands are with respect to collision. We estimated the expected number of collisions given the band positions, and described how our findings can be used to arrive at improved similarity coefficients for binary AFLP data.

Chapter 5: Homoplasy corrected estimation of genetic similarity from AFLP bands, and the effect of the number of bands on the precision of the estimate
In this paper, published in Theoretical and Applied Genetics, we proposed new estimators of genetic similarity using binary AFLP data. The new estimators are not hindered by the bias caused by homoplasy, that ordinary similarity coefficients like Dice and Jaccard suffer from. We also studied the precision of the estimators. As an application, we studied how the numbers of bands in the lanes affect the precision of the estimators. This has relevance for the design of AFLP experiments: how many bands per lane should we strive for?

Chapter 6: Codominant scoring of AFLP
This chapter describes the codominant scoring of AFLP in the case of diploid organisms, i.e. the genotype calling of bands into homozygous present (AA), heterozygous (Aa), or homozygous absent (aa), given their intensity. The methodology, already described by (R. C. Jansen, Geerlings, van Oeveren, & van Schaik, 2001) and (Piepho & Koch, 2000), is based on normal mixture models. We describe an application of the methodology in an association panel of tomato, using in-house developed software in R. The software contains some features that may enhance the unmixing of the distributions. We touch upon the relationship between codominant scoring and the problem of collision and homoplasy.

Chapter 7: Discussion
In the final chapter we summarize the results from the preceding chapters, and compare our quantitative results with the findings from literature, which we compiled in section 1.4. We discuss the relevance of our results for AFLP practice. We also sketch future work, showing some initial promising results.

Chapter 2

Significance Tests and Weighted Values for AFLP Similarities, Based on Arabidopsis *in Silico* AFLP Fragment Length Distributions [1]

by Wim J.M. Koopman and Gerrit Gort

2.1 Summary

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 approximately 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.

[1] Published as: W.J.M. Koopman and G. Gort (2004). Significance Tests and Weighted Values for AFLP Similarities, Bases on Arabidopsis *in Silico* AFLP Fragment Length Distributions. *Genetics*, 167:1915-1928

2.2 Introduction

AFLP is a DNA fingerprinting technique developed by Keygene N.V. (Vos et al., 1995). The technique consists of four steps: (1) digestion of DNA with two restriction enzymes, (2) ligation of double-stranded oligonucleotide adaptors 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 & LaReesa Wolfenbarger, 1999). In the 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 similiarity or dissimilarity coefficient, is used as a measure for genetic or phenetic relationship. Various coefficients have been developed to quantify (dis)similiarity, mainly differing in the weighting of comigrating relative no non-comigrating fragments (see, e.g. Nei & 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), Carduineae thistles (O'Hanlon & Peakall, 2000a), 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 *Glycyine max* (Meksem et al., 2001). The proportion of comigrating nonidentical fragments ranged from at least 10% within genotypes or among closely related (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 & Peakall, 2000a). Given the proportions of comigrating nonidentical bands, a serious overestimation of pairwise similarities among genotypes can be expected. Indeed, Karp, Seberg, and Buiatti (1996) noted that the occurrence of 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 fragment 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, Terauchi, Kahl, and Tajima (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 & 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.

2.3 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 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 of 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 assumptions 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 & 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 that the *G+C* contents of the majority of plants ranged between 35 and 50% (see, e.g. Marie & Brown, 1993; Barow & 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* possible occupies a special position with only 30% $G+C$ (Nagl & 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 adaptor sequences.

Figure 2.1 depicts the AFLP FLD for 35-50% $G+C$. The distributions show that the probability that a fragment will occur decreases with increasing fragment lengths 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 distribution: 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

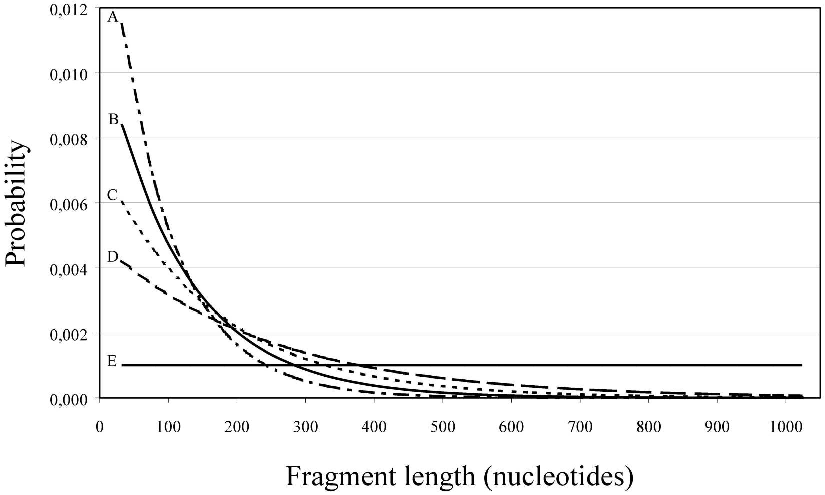


Figure 2.1: Theoretical AFLP FLDs based on Innan et al. (1999) for a genome with 35% $G+C$ (A), 40% $G+C$ (B), 45% $G+C$ (C), and 50% $G+C$ (D), respectively. The uniform distribution (E; equal probability for all fragments) is given as a reference.

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 2.2. 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 2.2 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 highest frequency in the Arabidopsis genome), and C/G (the nucleotides with the lowest

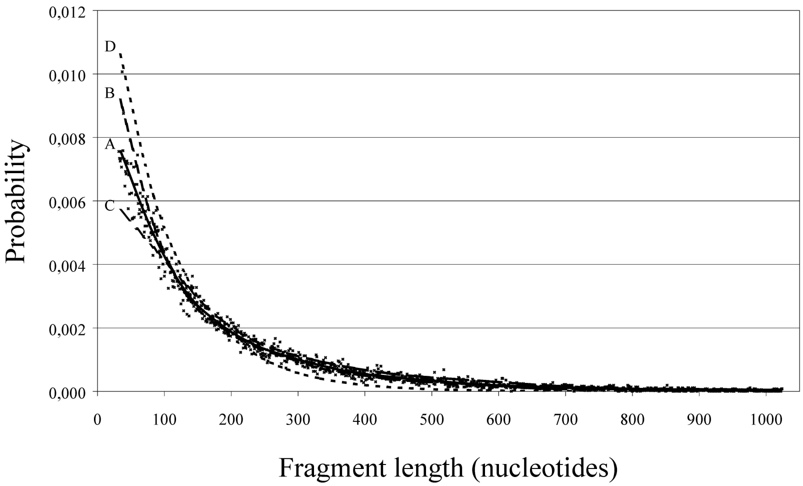


Figure 2.2: Relative frequency distribution of fragments resulting from *in silico* AFLP on the Arabidopsis genome sequence without selective nucleotides (frequencies for each length class are denoted by dots). (A) Smoothed FLD resulting from *in silico* AFLP on the Arabidopsis genome sequence without 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.

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 2.2 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 fragment (> 107 bp). The use of G/C 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 2.2 and 2.4, however, the AFLP FLDs appear as continuous 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 2.2 and 2.4 show both the smoothed discrete FLDs (line A in Figure 2.2; lines A and B in Figure 2.4) and the nonsmoothed discrete FLDs (probability in each length class depicted as 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 gen-

erated with the appropriate 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 patterns (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 number 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 & Li, 1979), while Jaccard (Jaccard, 1908) and simple matching similarity (Sokal & Sneath, 1963) 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 in 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 $= 1-$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 2.3 and Table 2.1 show results for the widely employed scoring range $50 - 500$ bp and an AFLP procedure with $A - C$ selective nucleotides. Figure 2.3 shows the relationship between the number of bands scored in each of two genotypes and the expected number of bands shared. Table 2.1 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 2.1, 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

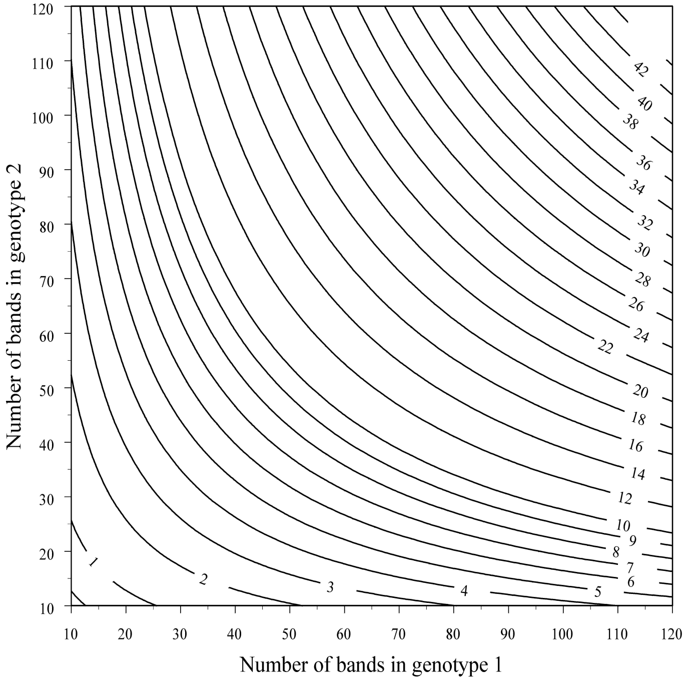


Figure 2.3: Relationship between number of bands scored in each of two genotypes and the expected number of bands shared. The lines depict whole numbers of expected shared bands; the actual numbers are inserted in the lines at the bottom and the right side of the plot. The plot corresponds to a scoring range of 50-500 bp and an AFLP procedure with A/C selective nucleotides.

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 2.1, 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 2.1) illustrates 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 2.1 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 numbers shows that such (dis)similarities, although low, may still be significant.

Table 2.1: Test statistics for scoring range 50-500 bp and an AFLP procedure with A/C selective nucleotides

n_1	n_2	Exp.bands	95%	99%	Exp.Dice	95%	99%	Max	Exp.Nei-Li	95%	99%	Min
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

(continued)

Table 2.1 (Continued)

n_1	n_2	Exp.bands	95%	99%	Exp.Dice	95%	99%	Max	Exp.Nei-Li	95%	99%	Min
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
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

n_1	n_2	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

(*continued*)

Table 2.1 (Continued)

n_1	n_2	Exp.Jaccard	95%	99%	Max	Exp.SM	95%	99%	Min	Max
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
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

Table 2.1: 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 with column 9 the minimum possible 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).

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 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 2.2), 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 genotypes (*e.g.*, 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 2.2 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 band has already been demonstrated for *Beta* and *Glycine 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 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 in 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 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 $p_i = 1 - [1 - p(\text{fragment at } i)]^{N_1}$, and $q_i = 1 - [1 - p(\text{fragment at } i)]^{N_2}$, where N_1 and N_2 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 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 argue 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

Table 2.2: 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 8 scoring ranges. Column 1: number of band 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.

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) 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 (43.3% $G + C$) 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 Biotech-

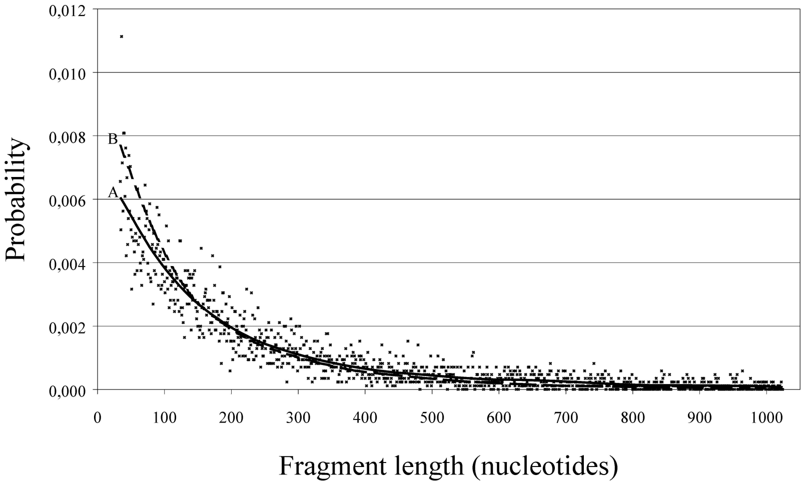


Figure 2.4: Relative frequency distribution of fragments resulting from *in silico* AFLP without selective nucleotides on the rice genome sequence (frequencies for each length class are denotes by dots). (A) Smoothed FLD resulting from *in silico* AFLP without selective nucleotides on the rice genome sequence. (B) The smoothed FLD resulting from *in silico* AFLP without selective nucleotides on the Arabidopsis genome sequence is given as a reference. Fragments lengths range from 32 to 1024 bp.

nology 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 2.4. 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, Salinas, Matassi, & Bernardi, 1990; Devos, Beales, Nagamura, & Sasaki, 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 test 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 statistcs 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.

2.4 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, Matassi, Montera, & Bernardi, 1988; Matassi, Montero, Salinas, & Bernardi, 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 Arabidopsiis *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 or 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 ALFP 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 2.1 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 $\sim 4\%$. For 30 bands, this proportion is $\sim 12\%$, for 60 bands it is $\sim 22\%$, for 90 bands it is $\sim 31\%$, and for 120 bands it is $\sim 40\%$. 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 2.1 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.

2.5 Acknowledgements

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.

Chapter 3

Fragment length distributions and collision probabilities for AFLP markers [1]

by Gerrit Gort, Wim J.M. Koopman and Alfred Stein

3.1 Summary

AFLP is a DNA fingerprinting technique frequently used in the plant and animal sciences. A drawback of the technique is the occurrence of multiple DNA fragments of the same length in a single AFLP lane, which we name a collision. In this paper we quantify the problem. The well-known birthday problem plays a role. Calculation of collision probabilities requires a fragment length distribution (fld). We discuss three ways to estimate the fld: based on theoretical considerations, on in-silico determination using DNA sequence data from *Arabidopsis thaliana*, or on direct estimation from AFLP data. In the latter case we use a generalized linear model with monotone smoothing of the fragment length probabilities. Collision probabilities are calculated from two perspectives, assuming known fragment counts and assuming known band counts. We compare results for a number of fld's, ranging from uniform to highly skewed. The conclusion is that collisions occur often, with higher probabilities for higher numbers of bands, for more skewed distributions and, to a lesser extent, for smaller scoring ranges. For a typical plant genome an AFLP with 19 bands is likely to contain the first collision. Practical implications of collisions are discussed. AFLP examples from lettuce and chicory are used for illustration.

[1] Published as: G. Gort, W.J.M. Koopman and A. Stein (2006). Fragment length distributions and collision probabilities for AFLP markers. *Biometrics*, 62:1107-1115

3.2 Introduction

AFLP is a DNA fingerprinting technique, developed by Keygene N.V. (Vos et al., 1995). AFLP fingerprints, also called profiles, are generated in four steps: 1) DNA is cut into fragments by two restriction enzymes. We focus on the enzymes *Mse*I and *Eco*RI. *Mse*I cuts the DNA strand at nucleotide sequence T-TAA, whereas *Eco*RI cuts at G-AATTC. 2) Adaptors are ligated to the fragments. 3) Selected fragments are amplified by the Polymerase Chain Reaction using primers complementary to the adaptors and restriction sites. The number of amplified fragments can be limited by using primers with more selective nucleotides. Each additional selective nucleotide will decrease the number of fragments approximately with a factor 4. The *Eco*RI primer is labelled with a radioisotope or fluorophore. 4) Fragments are separated by length on a gel on capillary electrophoresis system. The labelled fragments become visible as bands on the gel. A gel contains multiple lanes and within a lane the DNA fragments from one genome are sorted by length. Shorter fragments travel further in the gel, and therefore the position of a band in a lane corresponds to the length of the DNA fragments in the band. The length of the fragments in a band can be determined by comparing their position with the position of DNA fragments of known lengths (so-called sizers). Only fragments within a certain size domain, the scoring range, are observed, e.g. fragments with lengths 100-600 (yielding 501 possible band positions). AFLP bands are usually scored dominantly, meaning that a band is scored as either present or absent. From now on we will refer to both the AFLP technique and the AFLP fingerprint as “AFLP”. It will be clear from the context whether the technique or the fingerprint is meant.

AFLP’s can be generated for any organism, but the technique is most frequently used in the plant and animal sciences. It has become popular for a wide range of purposes, because it combines a high reproducibility with a high information content. AFLP’s are used for e.g. genome mapping and marker assisted breeding (M. J. W. Jeuken & Lindhout, 2004), phylogenetic studies (Koopman et al., 2001), conservation of plant genetic resources (McGregor, van Treuren, Hoekstra, & van Hintum, 2002) and identification of cultivars (Imazio et al., 2002).

Besides the advantages, AFLP’s also have a number of drawbacks. The most important drawback is the possible lack of homology of comigrating bands (Robinson & Harris, 1999). Homology of comigrating bands means that bands occurring at the same band position in different lanes indeed represent the same DNA fragment, originating from the same locus. Lack of homology of comigrating bands will be referred to as band size homoplasy. The effect of band size homoplasy on measures of association was examined in a previous study (Koopman & Gort, 2004). Closely related to band size homoplasy is comigration of different fragments of the same length within a single lane. With this type of comigration a band at a certain position comprises two or more fragments of the same length, originating from different loci. Comigration of different fragments within a single lane will be referred to as collision.

Collisions have been demonstrated in empirical data sets, mostly in a qualitative way. For example, Rouppe van der Voort et al. (1997) reported the recovery of multiple fragments from single AFLP bands “occasionally”. Mechanda et al. (2004)

conducted a detailed examination of the sequence identity of two AFLP markers in *Echinacea* and found multiple fragments. In a study on sugarbeet Hansen et al. (1999) reported that 13.2% out of 456 bands likely contained collisions. Hence, it seems that the occurrence of collisions is a general phenomenon and a possible serious source of error.

The aim of the present study is to quantify the occurrence of collisions from a theoretical point of view. A probabilistic description of AFLP is given in section 3.3. Collision probabilities are calculated based on fragment length distributions (fld's). In section 3.4 estimators of the fld are derived in three ways: in 3.4.1 from theoretical considerations, in 3.4.2 from sequence data and in 3.4.3 from empirical AFLP data. In section 3.5 the theory needed for the calculation of collision probabilities is described from two perspectives: in 3.5.1 from that of known fragment counts and in 3.5.2 (more realistically) from that of known band counts. Results for four different fld's are compared in section 3.6. In section 3.7 an example of AFLP data from a study on species relationships in lettuce and chicory (Koopman et al., 2001) is analysed. Results and biological implications of our findings are discussed in section 3.8.

3.3 Probabilistic description of AFLP

3.3.1 Basic assumptions, notation and probability model

We treat AFLP as a random sampling procedure of DNA fragments from a genome. Let m be the sample size, i.e. the number of amplified fragments. Since m is small compared to the total number of possible fragments in the genome, we treat the procedure as sampling with replacement.

A first representation of AFLP data is as fragment lengths $l_1, l_2, \ldots, l_m$. We assume that the lengths l_i are i.i.d. variables from a discrete fld F $(i = 1, \ldots, m)$. Let N be the number of band positions in the AFLP. A typical value of N is in the range 400-600. A value of $N = 501$ could arise e.g. when the scoring range is 100-600, or 50-550. We denote the lengths as $1, 2, \ldots, N$. Length 1 is the minimum length that can be scored on the gel. Denote with $p_j = P(l = j)$ $(j = 1, \ldots, N)$ the probability that a randomly drawn fragment has length l equal to j. Clearly, $\sum_{j=1}^{N} p_j = 1$. Notice that m and l_i are not directly observable and that not all l_i need to be different.

A second representation of AFLP data is as counts. Let k_j be the count of fragments of length j. Then, given m, the vector $k = (k_1, .., k_N)'$ has a multinomial distribution $Multinom(m, F)$. The k_j's are not directly observable.

We do observe bands at specific positions. We represent a band at position j $(j = 1, \ldots, N)$ as a binary variable y_j: $y_j = 0$ if the j^{th} position does not contain a band and $y_j = 1$ otherwise. Define the band probabilities $P_j = P(y_j = 1)$. Then $y_j \sim Bernoulli(P_j)$. We assume that the y_j's are independent, which seems reasonable if AFLP is treated as a random sampling procedure of fragments with replacement. The y_j's need not be identically distributed however, since P_j's generally are not the same for different j's. Notice that $y_j = 1$ means that *at least* 1 fragment of length j was formed. Therefore, the events $\{y_j = 1\}$ and $\{k_j \geq 1\}$ coincide.

Fragment length probabilities p_j and band probabilities P_j have the following relationship:

$$P_j = 1 - (1 - p_j)^m \tag{3.1}$$

because the probability of not finding a band at position j is the probability of no fragments at all of length j in a random sample of size m.
Let n be the number of bands in a lane, so $n = \sum_{j=1}^{N} y_j$. The number of collisions $z = m - n$. Notice that z has range 0 (no collisions) to $m - 1$ (all fragments equally long).

3.3.2 Occupancy and birthday problems for AFLP

The situation here described is well known in the field of Probability Theory. The number of bands n, given N, m and a uniform fld F, has an occupancy distribution (see Johnson, Kotz, & Kemp, 1992). The probability that $n < m$, i.e. the number of bands smaller than the number of fragments or at least one collision, is the probability of interest in the *birthday problem* (see e.g. Feller, 1968). The birthday problem, explained in AFLP terminology, states that the smallest number of fragments needed to have a collision probability of at least $\frac{1}{2}$ equals only $m = 23$ for an AFLP with $N = 365$ and a uniform fld F_U. A more realistic $N = 500$ results in $m = 27$.
The fld in AFLP's, however, is not uniform, but highly skewed (Koopman & Gort, 2004). The situation with a non-uniform F is called a generalized birthday problem, which is a recurrent topic of interest in probability theory, see e.g. Holst (1995) and Henze (1998). The probability of at least one collision will be at least as large as in the uniform case, as shown by Munford (1977). The distribution of n given N, m and a general fld F is called a generalized occupancy distribution (Chakraborty, 1993).

3.4 Fragment Length Distributions

Calculation of collision probabilities requires an estimate of the fld F. We obtain estimators in three different ways:

1. Use intrinsic properties of the AFLP procedure and simple assumptions about nucleotides to arrive at a theoretical (non-data driven) estimator F_T of F.

2. Use sequence data and simulation of the AFLP procedure to arrive at an "in-silico" estimator F_S of F.

3. Use empirical AFLP data to obtain an estimator F_A of F.

Estimation procedures 1 and a simpler version of 2 are described in Koopman and Gort (2004). We give an outline here, as they are at the basis of the present study.

3.4.1 Fld from theoretical assumptions

Innan et al. (1999) derived F_T based on theoretical considerations, assuming equal relative frequencies of the nucleotides $p_A = p_C = p_G = p_T = \frac{1}{4}$, which are constant across the genome and random nucleotide order. They find that

$$p_j = \frac{f^{(j-1)}(1-f)}{1-f^N} \tag{3.2}$$

where $f = (1 - p_A p_A p_T p_T)(1 - p_G p_A p_A p_T p_T p_C) = (1 - \frac{1}{4}^4)(1 - \frac{1}{4}^6) = 0.99585$; f is the probability that, walking along the DNA strand, no restriction site for *Mse*I or for *Eco*RI is found at the next nucleotide. The probabilities described by (3.2) form a truncated geometrical distribution. Truncation occurs, since no fragments longer than N are observable on the gel. We name the fld obtained in this way F_{T_1}.

Innan et al. (1999) assume equal relative frequencies of nucleotides. In real genomes however, the relative frequencies of A+T and G+C may be different, depending on the organism. The majority of plants have relative frequencies of G+C ranging from 0.35 to 0.50 (see e.g. Marie & Brown, 1993). E.g. the plant *Arabidopsis thaliana* has a GC content of 36%, resulting in $f = 0.98918$. We name the resulting fld F_{T_2}. F_{T_1} and F_{T_2} are shown in figure 3.1. Lower GC content will generally result in smaller fragments. This effect is caused mainly by the use of restriction enzyme *Mse*I, which will find more restriction sites, and therefore shorter fragment lengths, in GC-poor genomes.

3.4.2 Fld from sequence data

A second way of estimating the fld of a genome is to use available sequence data of a related organism and simulate the AFLP procedure using computer software. We used SAS (version 8.0) to this end. This “in silico” approach was applied to *Arabidopsis thaliana*, the entire genome sequence of which is available at The Institute for Genomic Research (TIGR) at http://www.tigr.org. Steps 1 and 2 of the AFLP procedure (restriction and adaptor ligation) were applied to the *Arabidopsis* genome. Only fragments with at least one *Eco*RI site were selected, rendering fragments containing numbers of nucleotides in the range 32-1024. In figure 3.1 the relative frequencies (left axis) and counts (right axis) for fragments with 51-700 nucleotides (lengths coded as 1-650) are shown, together with a histogram smoother F_S as estimator of the fld F. The histogram smoother comprises a generalized linear model (McCullagh & Nelder, 1989) using the Poisson distribution and log link for the counts and P-splines (Eilers & Marx, 1996). The penalty used in the P-splines was chosen at eyesight to arrive at a smooth fld. The effective dimension of the fit was about 6, with a residual Pearson's X^2 of 1150 on 644 df. This indicates that there is some overdispersion compared to Poisson variation. Investigation into the causes of the overdispersion revealed that some counts were extraordinary high and resulted often from centromeric regions of *A. thaliana*-chromosomes. We ignore the observed overdispersion, since it does not have a large impact on the estimate of the smooth fld F_S.

Comparison of the fld's F_{T_1}, F_{T_2} and F_S in figure 3.1 shows that F_S favors larger fragments than F_{T_2}, but smaller fragments than F_{T_1}. An obvious reason for the discrepancy between F_S and F_{T_2} (and F_{T_1}) is violation of underlying assumptions of F_T, like local variation in GC-content across the genome and non-randomness of the nucleotide order. On the other hand, F_S might be faulty as well, because we left out the third step of the AFLP procedure, which is the selective amplification using primers. If the nucleotide order is not random, the selection by a primer may disturb the original fld.

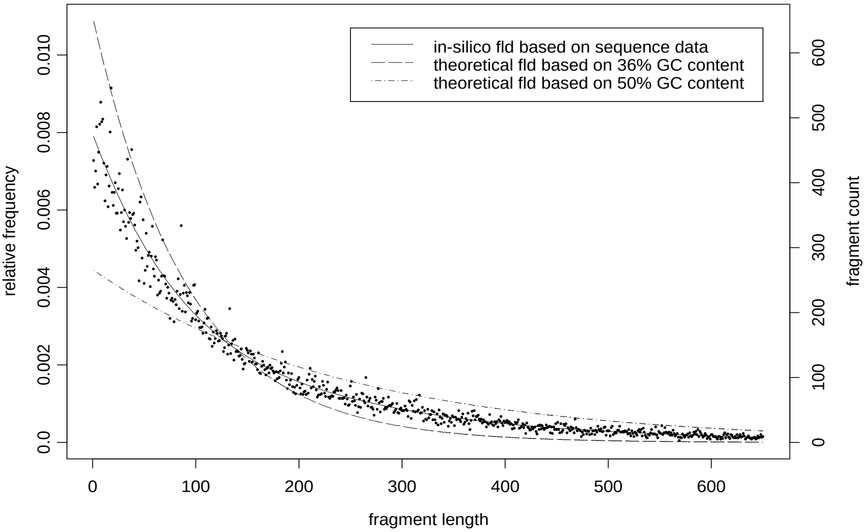


Figure 3.1: AFLP Fragment length distributions for *A. thaliana*. In-silico estimate F_S compared to theoretical estimates F_{T_2} and F_{T_1} based on 36% and 50% GC content

3.4.3 Fld from empirical AFLP data

Different species will generally have different fld's, depending e.g. on the GC content. Therefore, working with an estimate based on a possibly not too related species, e.g. the one from *A. thaliana*, is an oversimplification. Since sequence data are available for only a limited number of species, it is desirable to estimate F directly from the AFLP.

As described in section 3.3.1, AFLP data can be represented as a binary vector $y = (y_1, \dots, y_N)$ with $y_j \sim Bernoulli(P_j)$. From (3.1) it can be seen that

$$\log(-log(1 - P_j)) = log(-log((1 - p_j)^m) = log(m) + log(-log(1 - p_j)). \quad (3.3)$$

This suggests a generalized linear model for the binary vector y with complementary log-log (cll) link function and linear predictor $\eta_j = cll(P_j) = log(m) + cll(p_j)$.

Without restrictions this model is not useful, as we have as many parameters as binary observations. Therefore, we force the linear predictor η_j to change smoothly with the fragment length j. As in section 3.4.2 we use the P-spline approach. Motivated by the general shape of the fld's F_{T_1}, F_{T_2} and F_S, we enforce monotonicity of the smooth function, using an extra penalty (Bollaerts, Eilers, & van Mechelen, 2006). As result linear predictions $\hat{\eta_j}$ are obtained.
Rewriting the systematic part of the model into $cll(p_j) = cll(P_j) - log(m)$ and noting that $\sum_{j=1}^{N} p_j = 1$ we arrive at

$$\sum_{j=1}^{N} 1 - (e^{1/m})^{-e^{\eta_j}} = 1, \tag{3.4}$$

from which, for the linear predictions $\hat{\eta_j}$, m may be solved (e.g. using Newton-Raphson), rendering estimator $\hat{m}$. The resulting estimator F_{A_1} of F is given by

$$\hat{p_j} = 1 - e^{-e^{\hat{\eta_j} - log(\hat{m})}}. \tag{3.5}$$

A simpler route to estimator F_{A_2} is as follows. For small p_j (in the paper all $p_j < 0.011$), $-log(1 - p_j) \approx p_j$, so that (3.3) can be rewritten as $\eta_j = log(m) + log(-log(1 - p_j)) \approx log(m) + log(p_j) = log(mp_j)$. Therefore $e^{\eta_j} \approx mp_j$ and $\sum_{j=1}^{N} e^{\eta_j} = m \sum_{j=1}^{N} p_j = m$. The estimator F_{A_2} is given by

$$\hat{p_j} = e^{\hat{\eta_j}} / \sum_{i=1}^{N} e^{\hat{\eta_i}}. \tag{3.6}$$

Note that the same results follow by considering the unobserved fragment counts k_j, $Poisson(\lambda_j)$ distributed with expected counts λ_j and a log link function. Since $P_j = P(k_j > 0) = 1 - P(k_j = 0) = 1 - e^{-\lambda_j}$, it follows that $\eta_j = cll(P_j) = log(\lambda_j)$ (cf the derivation of the cll link in the dilution assay in McCullagh and Nelder (1989)). Estimator of p_j would be $e^{\hat{\eta_j}} / \sum_{i=1}^{N} e^{\hat{\eta_i}}$, equal to (3.6).
In a small simulation study with 10 randomly drawn AFLP's based on *A. thaliana*'s F_S with $m = 50$ we found negligible differences between F_{A_1} and F_{A_2}, and no systematic deviations of the estimates F_{A_1} and F_{A_2} from F_S. The estimates for p_1 (for F_S $p_1 = 0.0079$) were in the range 0.0056-0.0094, for p_{650} ($= 0.00013$) in the range $0.000032 - 0.00025$.

3.5 Collisions

In this section we focus on the occurrence of collisions. Since a band of an AFLP is supposed to represent a single DNA-fragment, the event that no collision occurs is of special importance. Theory is developed from two perspectives:

1. Fragment counts. The question how many fragments are allowed in order to have confidence in the absence of collisions has relevance for different aspects of the AFLP procedure, e.g. for the choice of restriction enzymes and primers,

and for the number of selective nucleotides. We are interested in the probability distribution of the collision count, given the fragment count. From this distribution the probability of no collision given the fragment count $P_0(m)$ and the smallest number of fragments m for which $P_0(m) < \frac{1}{2}$ are deduced in a straightforward way. Another quantity of interest is the expected number of fragments until the first collision occurs.

2. Band counts. In empirical AFLP datasets band counts are observed instead of fragment counts. Therefore, the probability distribution of the collision count, given the band count, is of interest. From it the probability of no collision given the band count $P_0(n)$ and the smallest number of bands for which $P_0(n) < \frac{1}{2}$ can be deduced.

3.5.1 Probability distribution of the number of collisions given the number of fragments

The required probability distribution is $P(z = z_0|m = m_0)$ with $z_0 = 0, \ldots, m_0 - 1$, or written as probabilities for the band count $P(n = m_0 - z_0|m = m_0)$. These probabilities form an occupancy distribution, as seen in section 3.3.2. The classical occupancy distribution refers to a situation with equal probabilities (uniform F_U), for which exact probabilities can be calculated easily. The generalized occupancy distribution with unequal probabilities is much less tractable.

Chakraborty (1993) derives exact formulae for all factorial moments $\mu_{[r]}$ of the generalized occupancy distribution. Because of computational limitations calculations in our situation are only feasible for the first 4 factorial moments.

Some specific occupancy probabilities are relatively easy to obtain. Take the probability of no collision given m: $P_0(m) = P(z = 0|m)$. From 3.3.1 we know that an AFLP with m fragments can be represented as a multinomial vector of counts $k = (k_1, \ldots, k_N) \sim Multinom(m, F)$, for which

$$P_0(m) = P(k_1 \leq 1, k_2 \leq 1, \ldots, k_N \leq 1 \mid \sum_j k_j = m) \tag{3.7}$$

This is a probability from the multinomial cumulative distribution function (cdf). For large N straightforward computation of the cdf probabilities is troublesome. In the literature a number of solutions is suggested. A remarkably good approximation of the cdf using Edgeworth expansions is given by Levin (1981), who writes the cdf probability as a product of terms that involves the convolution of N independent truncated Poisson variables. An improvement of his approximation is given by Butler and Sutton (1998), employing a saddlepoint approximation. A recursive calculation scheme resulting in the exact probability (3.7) is given by Sandell (1991). His approach is only applicable for probabilities that all k_i are less or equal to the same value j, being $j = 1$ in (3.7). We apply his recursive algorithm in section 3.6.1.

Calculation of $P_0(m)$ for increasing numbers of fragments m solves the generalized birthday problem. The smallest m for which $P_0(m) < \frac{1}{2}$ is the required number.

The probability that exactly one collision occurs $P(z = 1|m)$, can be calculated in the same way. The collision may have occurred at any of the N possible positions,

so we have to sum N multinomial probabilities $P(k_1 \leq 1, \ldots, k_j = 2, \ldots, k_N \leq 1 \mid \sum_j k_j = m)$, each of which may be approximated with a saddlepoint approximation. The situation becomes more difficult for higher values of z. For $z = 2$ there may be a collision at two different positions, or 2 collisions at a single position, requiring $\binom{N}{2} + N$ evaluations of approximated multinomial probabilities. This number becomes intractable for large numbers of collisions.
Holst (1995) studies, phrased in AFLP terminology, the expected number of fragments until a given number of collisions occurs. Application of his results to the situation of the first collision gives

$$E(m) = \int_0^\infty e^{-t} \prod_{i=1}^{N} (1 + p_i t)\, dt \tag{3.8}$$

which, using the gamma-function $\Gamma(i) = \int_0^\infty e^{-t} t^i dt = i!$ can be written as

$$E(m) = 1 + 1! \sum_i p_i + 2! \sum_{i \neq j} p_i p_j + 3! \sum_{i \neq j \neq k} p_i p_j p_k + \ldots + N! p_1 p_2 \ldots p_{N-1} p_N. \tag{3.9}$$

We developed a special calculation scheme to calculate (3.9).

3.5.2 Probability distribution of the number of collisions given the number of bands

The required probability distribution equals $P(z = z_0 | n = n_0) = P(m = n_0 + z_0 | n = n_0)$. With Bayes' rule (and $m_0 = n_0 + z_0$) we get

$$\begin{aligned} P(m = m_0 | n = n_0) &= \frac{P(n = n_0 | m = m_0) \times P(m = m_0)}{P(n = n_0)} = \\ &= \frac{P(n = n_0 | m = m_0) \times P(m = m_0)}{\sum_{i \geq n_0} P(n = n_0 | m = i) \times P(m = i)}. \end{aligned} \tag{3.10}$$

The prior distribution of m is unknown, as it depends for instance on the genome size and the number of selective nucleotides of the primers. If we assume a uniform distribution of m, which can be only approximately the case, (3.10) simplifies into

$$P(m = m_0 | n = n_0) = \frac{P(n = n_0 | m = m_0)}{\sum_{i \geq n_0} P(n = n_0 | m = i)}. \tag{3.11}$$

The probabilities making up the right hand side of (3.11) stem from the generalized occupancy distribution. Because this distribution is intractable in general (see 3.5.1), we approximate it and calculate probabilities (3.11) using the approximating distribution. A promising paper by Kathman and Terrell (2003), suggesting a Poisson approximation by constrained exponential tilting, did not give useful results. Instead we use a simple binomial approximation by equating the first two moments of the distribution of z (see section 3.5.1) and the binomial distribution, resulting in values for the binomial parameters p and (possibly non-integer) n. This procedure gives acceptable approximations for values of z

not too far in the tail. As an example take F_U with $N = 500$ and $m = 20$, giving first 2 central moments of z 0.375 and 0.357. The approximating binomial distribution has parameters $n = 7.688$ and $p = 0.0488$. Exact probabilities are $P(z = 0) = 0.6804, P(z = 1) = 0.2688, P(z = 2) = 0.04600$, whereas the binomial approximation gives resp. 0.6805, 0.2686 and 0.04612. The approximation deteriorates in the tail: exact $P(z = 6) = 3.357 \times 10^{-7}, P(z = 7) = 6.679 \times 10^{-9}$ and approximated $P(z = 6) = 2.354 \times 10^{-7}, P(z = 7) = 2.915 \times 10^{-9}$. Better approximations, incorporating the information from the higher moments, should be possible, but for this study the binomial distribution seems good enough for calculation of expectation and standard deviation of z given n.

Using the same approach we can approximate the probability of no collision given n $P_0(n)$ and find the minimum number of bands n for which $P_0(n) < \frac{1}{2}$.

3.6 Results for collision probabilities, comparing fld's based on theoretical considerations and estimated from *A. thaliana*

In all calculations the fld is assumed to be known. We compare results for different fld's: 1)F_U as reference, 2) F_{T_1}, 3) F_{T_2} and 4) F_S (with scoring range starting at 51). Comparison of the results gives an indication of sensitivity for differences in fld's. We use numbers of band positions $N = 400$, 500 and 600, as these are common values. Furthermore, results are produced for fragment counts or band counts 10, 20,. . . , 120.

3.6.1 Probability distribution of the number of collisions given the number of fragments

Table 3.1 gives expectations and standard deviations of the number of collisions given the number of fragments and the probabilities of no collision $P_0(m)$. The fld F_U gives a lower bound for the expectation and upper bound for $P_0(m)$. Observe that: 1) larger fragment counts lead to more, and more likely collisions; 2) larger values of N result in only slightly less, and less likely collisions; 3) more skewed fld's have more, and more likely collisions. Clearly, fragments from such distributions tend to concentrate at smaller lengths and therefore have more collisions.

Results for the birthday problem for fragments with $N = 500$ and F_U, F_{T_1}, F_{T_2} and F_S are 27, 24, 17 and 20, respectively. So, for a realistic situation (F_S and $N = 500$) only $m = 20$ fragments are needed to have a likely collision ($P_0(m) < \frac{1}{2}$). For $N = 500$ the expected numbers of fragments until the first collision according to Holst (1995) for F_U, F_{T_1}, F_{T_2} and F_S are 28.7, 25.0, 17.8 and 21.0.

3.6.2 Probability distribution of the number of collisions given the number of bands

Table 3.2 gives expected numbers of collisions, standard deviations and probabilities of no collision given the number of bands $P_0(n)$. The values for F_U, serving as lower bounds for expectations, are based on exact occupancy probabilities, the

	N=400				N=500				N=600			
m	F_U	F_{T_1}	F_{T_2}	F_S	F_U	F_{T_1}	F_{T_2}	F_S	F_U	F_{T_1}	F_{T_2}	F_S
Expected number of collisions												
10	0.11	0.14	0.25	0.19	0.09	0.12	0.24	0.17	0.07	0.11	0.24	0.17
20	0.47	0.57	1.02	0.77	0.38	0.50	1.00	0.72	0.31	0.46	0.99	0.69
30	1.06	1.28	2.27	1.73	0.85	1.13	2.23	1.61	0.71	1.04	2.22	1.54
40	1.89	2.28	3.98	3.05	1.52	2.00	3.91	2.84	1.27	1.84	3.89	2.72
50	2.94	3.54	6.11	4.71	2.37	3.11	6.01	4.39	1.99	2.86	5.98	4.20
60	4.22	5.05	8.63	6.69	3.41	4.45	8.49	6.23	2.86	4.10	8.45	5.98
70	5.71	6.82	11.51	8.97	4.62	6.01	11.33	8.37	3.88	5.54	11.27	8.03
80	7.41	8.82	14.74	11.55	6.00	7.79	14.51	10.77	5.05	7.18	14.44	10.34
90	9.32	11.06	18.29	14.40	7.56	9.77	18.01	13.44	6.36	9.01	17.91	12.90
100	11.42	13.52	22.13	17.50	9.28	11.96	21.80	16.35	7.82	11.03	21.68	15.70
110	13.72	16.19	26.26	20.86	11.17	14.34	25.86	19.50	9.42	13.23	25.73	18.73
120	16.22	19.08	30.64	24.45	13.22	16.91	30.18	22.87	11.16	15.61	30.03	21.97
Standard deviation of number of collisions												
10	0.33	0.36	0.48	0.42	0.30	0.34	0.48	0.41	0.27	0.33	0.48	0.40
20	0.66	0.73	0.95	0.84	0.60	0.68	0.94	0.81	0.55	0.66	0.94	0.80
30	0.98	1.07	1.38	1.23	0.89	1.01	1.37	1.20	0.82	0.98	1.37	1.17
40	1.29	1.40	1.77	1.60	1.17	1.33	1.76	1.56	1.08	1.28	1.76	1.53
50	1.58	1.72	2.14	1.94	1.44	1.63	2.13	1.89	1.34	1.58	2.12	1.87
60	1.86	2.02	2.47	2.26	1.71	1.92	2.46	2.21	1.58	1.86	2.46	2.18
70	2.13	2.30	2.77	2.57	1.96	2.20	2.77	2.52	1.82	2.14	2.76	2.49
80	2.39	2.57	3.06	2.85	2.21	2.47	3.05	2.80	2.06	2.40	3.05	2.77
90	2.63	2.82	3.32	3.12	2.44	2.72	3.31	3.07	2.28	2.65	3.31	3.04
100	2.87	3.07	3.56	3.37	2.67	2.96	3.56	3.33	2.51	2.89	3.55	3.30
110	3.09	3.30	3.78	3.60	2.89	3.20	3.78	3.57	2.72	3.13	3.78	3.54
120	3.30	3.52	3.98	3.82	3.10	3.42	3.99	3.80	2.93	3.35	3.99	3.77
Probability of no collision												
10	0.89	0.87	0.78	0.83	0.91	0.89	0.78	0.84	0.93	0.90	0.78	0.85
20	0.62	0.56	0.34	0.45	0.68	0.60	0.35	0.47	0.73	0.63	0.35	0.49
30	0.32	0.26	0.084	0.16	0.41	0.31	0.088	0.18	0.48	0.34	0.089	0.20
40	0.13	0.086	0.011	0.035	0.20	0.12	0.012	0.046	0.27	0.14	0.013	0.053
50	0.041	0.021	< 0.001	0.005	0.079	0.035	< 0.001	0.008	0.12	0.046	0.001	0.010
60	0.009	0.004	< 0.001	< 0.001	0.025	0.008	< 0.001	< 0.001	0.047	0.012	< 0.001	0.001

Table 3.1: Expectation and standard deviation of the number of collisions, and probabilities of no collision given the fragment count

others on binomial approximations. In line with the results in 3.6.1 we observe that: 1) larger band counts lead to more, and more likely collisions; 2) the number of band positions N has only a mild influence; 3) more skewed fld's have more, and more likely collisions.

Comparison of tables 3.1 and 3.2 shows that probabilities of no collision given n are, obviously, smaller than given m. Differences in expected collision counts are small for small counts m and n, but huge for larger counts. E.g. for $N = 500$ and F_S we find for $m = 120$ 23 collisions but for $n = 120$ 47 collisions. Also notice the much larger standard deviations in the latter table.

Results for the birthday problem for bands with $N = 500$ and F_U, F_{T_1}, F_{T_2} and F_S are 26, 23, 16 and 19 respectively. So, for a realistic situation (F_S and $N = 500$) only $n = 19$ bands are needed to have a likely collision ($P_0(n) < \frac{1}{2}$).

	N=400				N=500				N=600			
n	F_U	F_{T_1}	F_{T_2}	F_S	F_U	F_{T_1}	F_{T_2}	F_S	F_U	F_{T_1}	F_{T_2}	F_S
Expected number of collisions												
10	0.14	0.17	0.32	0.24	0.11	0.15	0.31	0.22	0.09	0.14	0.31	0.21
20	0.54	0.67	1.26	0.93	0.43	0.58	1.24	0.86	0.36	0.53	1.23	0.82
30	1.23	1.51	2.88	2.11	0.97	1.31	2.83	1.94	0.80	1.19	2.81	1.84
40	2.20	2.71	5.27	3.81	1.73	2.35	5.16	3.50	1.43	2.14	5.13	3.33
50	3.48	4.31	8.50	6.08	2.74	3.72	8.32	5.57	2.25	3.38	8.26	5.29
60	5.10	6.32	12.67	8.97	3.98	5.43	12.38	8.19	3.27	4.93	12.29	7.77
70	7.05	8.77	17.89	12.53	5.49	7.51	17.47	11.40	4.50	6.80	17.33	10.79
80	9.38	11.69	24.30	16.81	7.27	9.97	23.69	15.24	5.94	9.01	23.49	14.40
90	12.10	15.12	32.05	21.88	9.34	12.85	31.19	19.76	7.60	11.58	30.90	18.63
100	15.24	19.10	41.30	27.81	11.70	16.15	40.12	25.02	9.49	14.52	39.73	23.53
110	18.82	23.66	52.27	34.69	14.37	19.92	50.67	31.07	11.63	17.87	50.15	29.15
120	22.88	28.86	65.19	42.60	17.38	24.17	63.06	37.97	14.01	21.62	62.36	35.55
Standard deviation of number of collisions												
10	0.38	0.42	0.58	0.50	0.34	0.39	0.57	0.48	0.31	0.37	0.57	0.47
20	0.75	0.84	1.18	1.00	0.67	0.78	1.17	0.96	0.61	0.74	1.17	0.94
30	1.14	1.27	1.84	1.54	1.01	1.18	1.82	1.47	0.91	1.13	1.81	1.44
40	1.54	1.73	2.56	2.11	1.35	1.60	2.53	2.02	1.22	1.53	2.52	1.96
50	1.95	2.21	3.35	2.73	1.71	2.04	3.31	2.60	1.55	1.94	3.29	2.52
60	2.38	2.71	4.22	3.38	2.08	2.50	4.16	3.21	1.87	2.37	4.14	3.12
70	2.83	3.24	5.18	4.09	2.47	2.97	5.10	3.87	2.21	2.82	5.08	3.75
80	3.30	3.80	6.24	4.85	2.86	3.47	6.14	4.57	2.56	3.28	6.11	4.42
90	3.79	4.39	7.43	5.67	3.27	3.99	7.30	5.32	2.91	3.76	7.26	5.14
100	4.30	5.00	8.76	6.55	3.69	4.53	8.59	6.13	3.28	4.27	8.53	5.90
110	4.83	5.66	10.25	7.50	4.12	5.11	10.03	6.99	3.65	4.79	9.96	6.71
120	5.39	6.35	11.92	8.54	4.57	5.70	11.64	7.91	4.03	5.34	11.55	7.58
Probability of no collision												
10	0.87	0.84	0.73	0.79	0.90	0.86	0.74	0.81	0.91	0.87	0.74	0.81
20	0.59	0.52	0.30	0.41	0.65	0.57	0.31	0.44	0.70	0.59	0.31	0.45
30	0.30	0.23	0.070	0.14	0.39	0.28	0.074	0.16	0.46	0.32	0.075	0.17
40	0.12	0.076	0.009	0.030	0.19	0.11	0.010	0.039	0.25	0.13	0.010	0.046
50	0.036	0.018	< 0.001	0.004	0.071	0.030	< 0.001	0.006	0.11	0.041	< 0.001	0.008
60	0.008	0.003	< 0.001	< 0.001	0.022	0.006	< 0.001	< 0.001	0.043	0.010	< 0.001	0.001

Table 3.2: Approximated expectation and standard deviation of the number of collisions, and probabilities of no collision given the band count

3.7 AFLP examples on lettuce and chicory

In a study on species relationships in lettuce Koopman et al. (2001) generated AFLP's (N = 392, shortest fragment length 110) on lettuce and chicory. We selected 4 species from this study to cover a wide range of band counts: *Lactuca tenerrima* with 10 individual plants (n in the range 28 − 33), *Cichorium intybus* with 5 plants (n in the range 37 − 40), *L. sativa* with 11 plants (n in the range 47 − 56) and *L. tatarica* with 12 plants (n in the range 75 − 100). Taxonomically the selected plants are close to *A. thaliana*: all are angiosperms with GC contents in the range 36-38%. Therefore, F_S is expected to be a reasonable fld. In figure 3.2 the fld's F_{A_1} estimated from the individual AFLP's are plotted. F_S is plotted as a reference. The estimated 11 fld's for *L. sativa* resemble F_S well. There are larger differences for the estimated fld's of *L. tenerrima* (running flatter), *L. tatarica* (running flatter) and *C. intybus* (running steeper). The variability of the fld's

Species	n	F_U	F_{T_1}	F_{T_2}	F_S	F_{A_1}
L. tenerrima	28	1.09 (1.07)	1.33 (1.19)	2.51 (1.71)	1.70 (1.37)	1.31 (1.18)
C. intybus	40	2.25 (1.55)	2.75 (1.74)	5.28 (2.56)	3.55 (2.02)	5.18 (2.53)
L. sativa	56	4.51 (2.24)	5.54 (2.53)	10.91 (3.86)	7.18 (2.97)	7.63 (3.08)
L. tatarica	100	15.62 (4.36)	19.43 (5.06)	41.45 (8.76)	25.65 (6.17)	22.33 (5.57)

Table 3.3: Approximated expected number of collisions (s.d.) given the band count for four AFLP profiles from *Lactuca* and *Cichorium*

within a species is comparable to the variability found in a small simulation study on randomly drawn AFLP's from F_S (see section 3.4.3). This suggests that the fld is constant for a species and can better be estimated from all available AFLP information of plants of that species.

For each of the four species used in figure 3.2 we choose the most extreme individual plant with respect to the number of bands in its AFLP: for *L. tenerrima* a plant with 28 bands, for *C. intybus* a plant with 40 bands, *L. sativa* a plant with 56 bands and *L. tatarica* a plant with 100 bands. Next, we calculated the approximate expectations and standard deviations of the collision count given the band count for each of the 4 plants, based on the fld's F_U, F_{T_1}, F_{T_2}, F_S and F_{A_1}, estimated from the individual AFLP profiles. The results are given in table 3.3. We expect collision counts based on F_{A_1} close to F_S as motivated above. Note that the effect of type of fld is relatively mild for e.g. *L. tenerrima* with only 28 bands (1-2.5 collisions are expected), but large for *L. tatarica* with 100 bands (19 collisions for F_{T_1} but 41 for F_{T_2}). The results for F_{A_1} are in the range given by F_{T_1} and F_{T_2} and not far from F_S.

3.8 Conclusions and discussion

As a basis for the calculation of collision probabilities we examined different estimators of AFLP fld's: F_{T_1}, F_{T_2}, F_S and, if available, F_{A_1}. The uniform F_U was included as a reference. The general shape is similar for all four fld's. Our findings confirm the conclusions of Innan et al. (1999) and Vekemans et al. (2002) that 1) the fld's are highly asymmetrical, 2) shorter fragments are more abundant than longer ones, and 3) the fld's vary with GC content of the genome. For the 0.36 GC content, typical for the *A. thaliana* genome, F_{T_2} based on Innan et al. (1999) shows an excess of short fragments compared to the in-silico estimate F_S. Estimator F_{A_1} for plants with GC contents not too different from *A. thaliana*'s gave results which were comparable to F_S, indicating the utility of F_S as standard fld for plants which are related to *A. thaliana*. The fld's F_{T_1} and F_{T_2} can be improved by allowing for dinucleotide frequencies and dependence between the occurrences of an TTAA- and GAATTC-restriction site. This is a point for further study.

Collision probabilities were calculated following two approaches: given the fragment count and given the band count. In general, more skewed fld's and higher numbers of fragments or bands result in more, and more likely collisions. The number of band positions has only a mild effect (depending on the skewness of the fld). Focusing on F_S with 500 band positions only 20 fragments (19 bands)

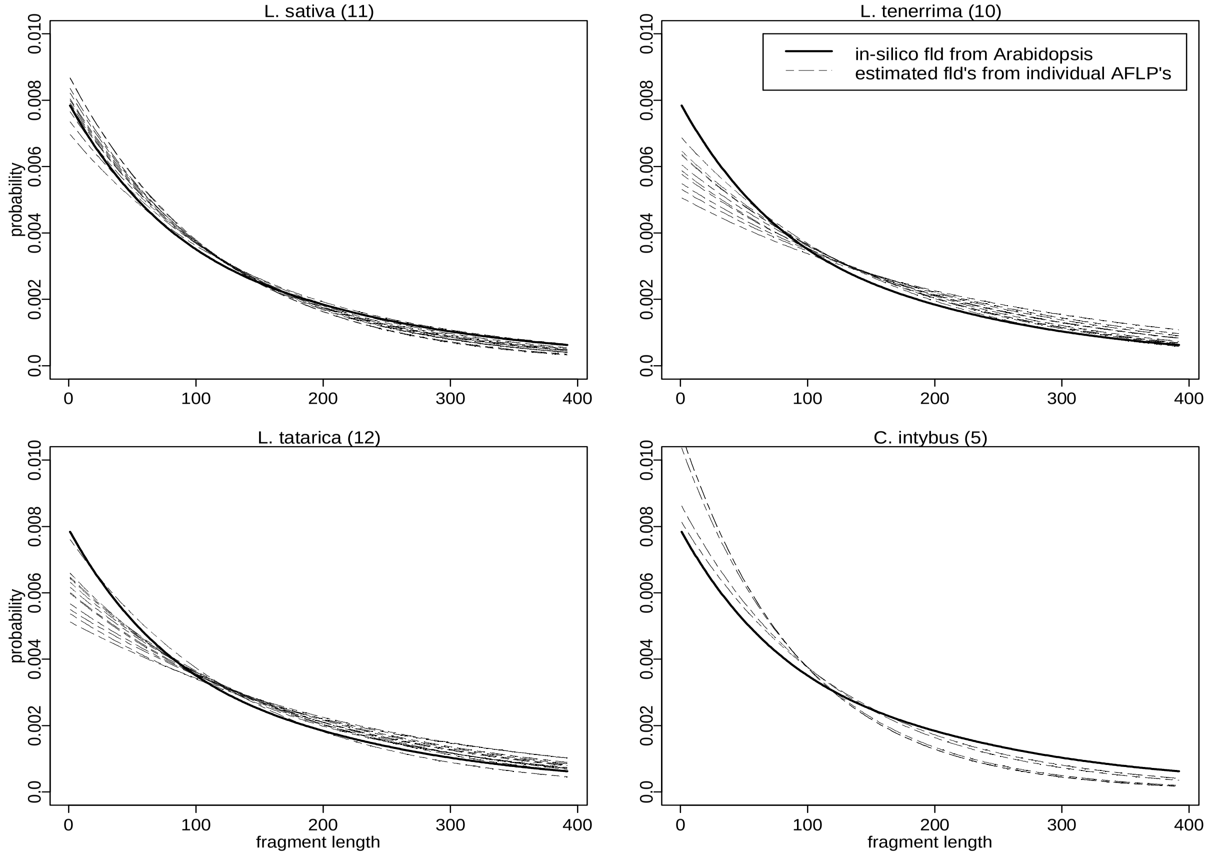


Figure 3.2: Estimated AFLP fragment length distributions F_{A_1} of some *Lactuca* and *Cichorium* species based on glm's with monotone P-splines; numbers in brackets denote number of individual AFLP profiles, i.e. plant individuals

already result in a collision probability above 0.5. The first collision is expected in AFLP's with 21 fragments.
The expected number of collisions rapidly increases with the number of fragments or bands. For 20 bands it is 0.8(±0.5) , so one of the bands is expected to contain multiple fragments (4%). For 50 bands, 5.5(±2.6) bands are expected to contain multiple fragments (12%), whereas for 100 bands this increases to 24(±6) (24%). These results are in line with an empirical study on sugarbeet by Hansen et al. (1999), who scored 456 bands in 16 AFLP lanes, giving an average of 28.5 bands per lane. They report that 13.2% of the bands were likely to contain collisions, whereas our results suggest $\approx 1.8 \pm 1.4$ collisions, giving a rough upper limit of 16%.
The present study indicates that collisions are a general phenomenon in AFLP's. The most important conclusion from our study is that, in order to avoid collisions, the number of fragments or bands in an AFLP profile must be much lower than the 50 to 100 originally recommended by Vos et al. (1995). Even band numbers as low as 20 are no guarantee for the absence of collisions.
Smaller numbers of bands are undesirable in the sense that they reduce the information content of the AFLP profile. Therefore it may be worthwhile to examine in some applications of AFLP, e.g. diversity studies, approaches to optimize the signal to noise ratio.
An alternative way to increase the information content of AFLP's is the codominant scoring of bands (Piepho & Koch, 2000). Here our findings have implications as well. In codominant scoring efforts are made to infer heterozygosity or homozygosity from the intensity of a band. Higher band intensities can be explained by collisions as well as by homozygosity (or other disturbing factors as PCR-effectiveness). Better inference on zygosity may be achieved by inclusion of fragment length information, since shorter fragments are generally more abundant and will likely show more collisions. In that respect shorter fragments have lower data quality. Inclusion of fragment length information, either through weights or by improved co-dominant scores, may result e.g. in better estimation of heterozygosity, testing of Hardy-Weinberg Equilibrium and mapping of QTL's. Papers about the effect of the band position and about codominant scoring are in preparation.

3.9 Acknowledgements

We are grateful to Terry Speed from the Walter and Eliza Hall Institute of Medical Research in Melbourne for useful discussion and a fruitful sabbatical period in Melbourne, and to Paul Eilers from the Leiden University Medical Center for providing software for monotonic spline smoothing. Also thanks to the anonymous reviewers whose comments improved the article.

Chapter 4

Collision probabilities for AFLP bands, with an application to simple measures of genetic similarity[1]

by Gerrit Gort, Wim J.M. Koopman, Alfred Stein, and Fred van Eeuwijk

4.1 Summary

AFLP is a frequently used DNA fingerprinting technique that is popular in the plant sciences. A problem encountered in the interpretation and comparison of individual plant profiles, consisting of band presence - absence patterns, is that multiple DNA fragments of the same length can be generated, that eventually show up as single bands on a gel. The phenomenon of two or more fragments coinciding in a band within an individual profile is a type of homoplasy, that we call collision. Homoplasy biases estimates of genetic similarity. In this study, we show how to calculate collision probabilities for bands as a function of band length, given the fragment count, the band count, or band lengths. We also determine probabilities of higher order collisions, and estimate the total number of collisions for a profile. Since short fragments occur more often, short bands are more likely to contain collisions. For a typical plant genome and AFLP procedure, the collision probability for the shortest band is 25 times larger than for the longest. In a profile with 100 bands a quarter of the bands may contain collisions, concentrated at the shorter band lengths. All calculations require a careful estimate of the monotonically decreasing fragment length distribution. Modifications of Dice and Jaccard coefficients are proposed. The principles are illustrated on data from a phylogenetic study in lettuce.

[1] Published as: G. Gort, W.J.M. Koopman, A. Stein, and F.A. van Eeuwijk. Collision probabilities for AFLP bands, with an application to simple measures of genetic similarity (2008). *Journal of Agricultural Biological and Environmental Statistics*, 13(2):1-22

4.2 Introduction

4.2.1 AFLP

Amplified Fragment Length Polymorphism (AFLP) is a DNA fingerprinting technique, widely used in the plant sciences (e.g. M. R. Jeuken, Van Wijk, Peleman, & Lindhout, 2001), and to a lesser extent in the animal (e.g. Duim et al., 2001) and human sciences (e.g. Prochazka et al., 2001). To understand the problem addressed in this paper, some insight into the molecular principles of AFLP is needed. In AFLP a sample of DNA fragments from a genome is produced in four steps. First, the DNA strands are cut into fragments using two restriction enzymes. Next, adaptors are ligated to the ends of the fragments. The adaptors allow primers to attach to the fragments. In the third step a selection of fragments is amplified using primers with selective nucleotides. The number of selective nucleotides determines the proportion of fragments to be amplified, each extra nucleotide roughly yielding a fourfold reduction. In the last step electrophoresis is used to separate the fragments by length, causing shorter fragments to travel further within a lane of a gel or capillary system. The amplified fragments are visualized as bands. The result for a genome is a kind of bar code of bands, called a profile. For a detailed description of the technique see Vos et al. (1995). Figure 4.1 shows part of an AFLP gel from a phylogenetic study on lettuce by Koopman et al. (2001), referred to as "the lettuce study".

The position of a band in a profile depends on the fragment length and is called band length. Only bands within a certain size domain are scored. In the lettuce study, only bands corresponding to fragments with 111-502 basepairs were taken into consideration. Usually bands are scored as present or absent, so-called dominant scoring. A profile can then be represented as a binary vector.

AFLP has a high replicability, and is able to produce large numbers of highly polymorphic bands (Mueller & LaReesa Wolfenbarger, 1999). However, it has a number of drawbacks. An important drawback in AFLP is size homoplasy (Robinson & Harris, 1999): comigrating bands in two different profiles are not necessarily homologous, that is, two bands with the same length may represent different DNA fragments, originating from different loci. The problem is well recognized (e.g. O'Hanlon & Peakall, 2000b). Mechanda et al. (2004) performed an extensive case study on comigrating bands, sequencing one monomorphic and one polymorphic band for plants from different taxonomic levels within the genus *Echinacea*. They found 59% sequence similarity within the genus for the monomorphic band, and as low as 1% for the polymorphic band. Althoff et al. (2007) studied homoplasy by electronically simulating AFLP on genomes from eight organisms, and found increasing homoplasy for increasing numbers of fragments. Koopman and Gort (2004) studied the effect of band size homoplasy on measures of similarity.

Another type of size homoplasy is comigration of fragments within a profile. Since fragments are separated by length, two or more fragments of the same length but from different loci with different nucleotide composition will appear as a single band. To discriminate the second type of size homoplasy from the first, we call the occurrence of multiple different fragments of the same length within a profile a collision. Hansen et al. (1999) reported in a study on sugar beet that 60 out of of 456 bands (13%) likely contained collisions. Vekemans et al. (2002) did a more

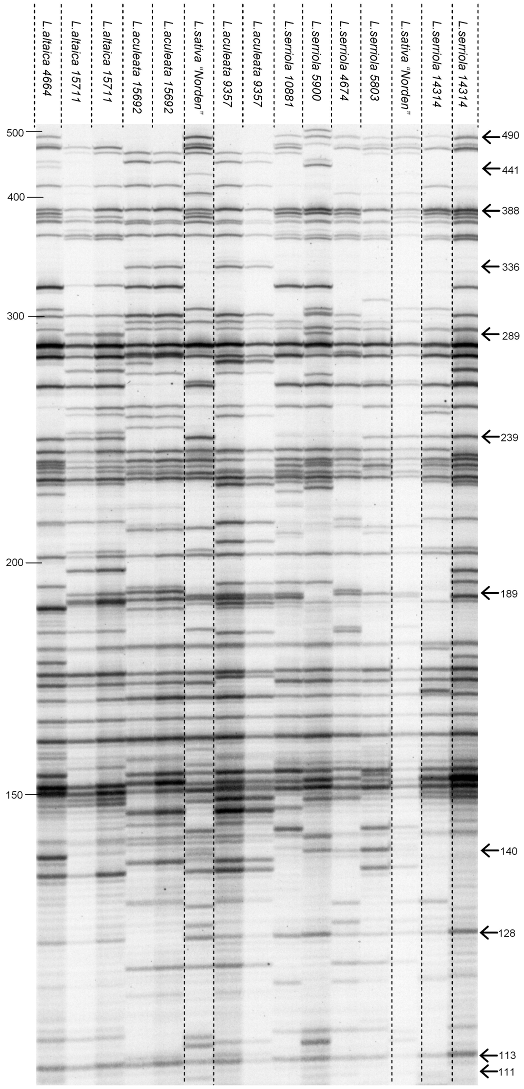


Figure 4.1: Part of an AFLP gel from a study on lettuce and related genera. Each of the 15 profiles represents a genome. The reference genotype *L. sativa* "Norden" occurs twice. The numbers behind the species names indicate accession numbers of the Centre for Genetic Resources, The Netherlands. The horizontal line segments are the AFLP bands with scoring range of 111-502 base pairs. Fragment lengths for a selection of bands are denoted to the right of the gel. The profiles, delimited with dashed lines, are used in further analysis.

general study, observing a relationship between degree of homoplasy and fragment size.

4.2.2 Motivating example

In the analysis of AFLP data, collisions and band size homoplasy are largely ignored. Here we focus on one of the applications of AFLP data: the estimation of genetic similarity between genotypes, e.g. in relationship studies like the lettuce study (Koopman et al., 2001), and in essential derivation (Eeuwijk & Law, 2004). In these studies simple pairwise similarity coefficients for binary data serve as estimators of genetic similarity between two genotypes (Kosman & Leonard, 2005). Usually a number of primer pairs is employed, resulting in multiple gels, and multiple profiles per genotype. In the lettuce study 95 accessions from 20 species of lettuce and related genera were fingerprinted with two primer combinations. From the binary scores Dice (D) and Jaccard (J) similarity coefficients were calculated. The coefficients were analysed with cluster analysis and principal coordinate analysis to study the species relationships.

D and J are defined as follows. For any pair of genotypes, a is the total number of shared bands from all profiles, b the number of bands occurring in the profiles of the first, but not of the second genotype, and c the number of bands occurring in the profiles of the second, but not of the first genotype. D is defined as $\frac{2a}{2a+b+c}$. $J = \frac{a}{a+b+c}$, and is directly related to D, since $J = D/(2-D)$. We define the genetic similarity S between two genotypes as the fraction of the genome that the two genotypes have in common, assuming for simplicity that the two genomes are equally sized. Now also a fraction S of the AFLP fragments is expected to be shared. D is supposed to estimate S, whereas J estimates $S/(2-S)$. This can be seen from an example: assume that two individuals share half of their equally sized genomes, so $S = \frac{1}{2}$. Then also half of the AFLP fragments are expected to be shared, so in expectation $a = b = c$, and Dice $D = \frac{2a}{2a+a+a} = \frac{1}{2}$, whereas Jaccard $J = \frac{a}{a+a+a} = \frac{1}{3}$. The Dice similarity coefficient has a natural interpretation in the genetic context, as we define it here, and is therefore to be preferred. In the following we will consider AFLP profiles generated by a single primer combination. Note that in most practical studies multiple primer combinations are used, and the information from the resulting multiple profiles is combined.

To appreciate the possible consequences of ignoring size homoplasy and collision in studies like the lettuce study, we performed a small simulation study. We sampled pairs of related profiles for given band counts (range 1-150), and for given genetic similarities $S = 0, 0.25, 0.50, 0.75$, and 0.90. Each pair was sampled in two steps. First, fragments were sampled for the first profile until a given number of bands was reached. In the second step, each fragment from the first profile had probability S to occur in the second, and new fragments were added until the given number of bands was reached for the second profile. Figure 4.2 shows averages of D and J, calculated from the binary band scores of the simulated pairs of profiles. Both D and J seriously overestimate the true similarity, with increasing bias for larger band counts and for smaller similarities. E.g. at true similarity $S = 0$, the biases of the average Dice coefficients are 0.18, 0.33, and 0.46 for band counts 50, 100, and 150 respectively. At true similarity $S = 0.50$ these values are 0.071, 0.14, and 0.20, and at S = 0.90 0.011, 0.022, and 0.032. Biases are caused both by comigrating fragments between profiles and by collision. In practice band counts larger than 100 are uncommon.

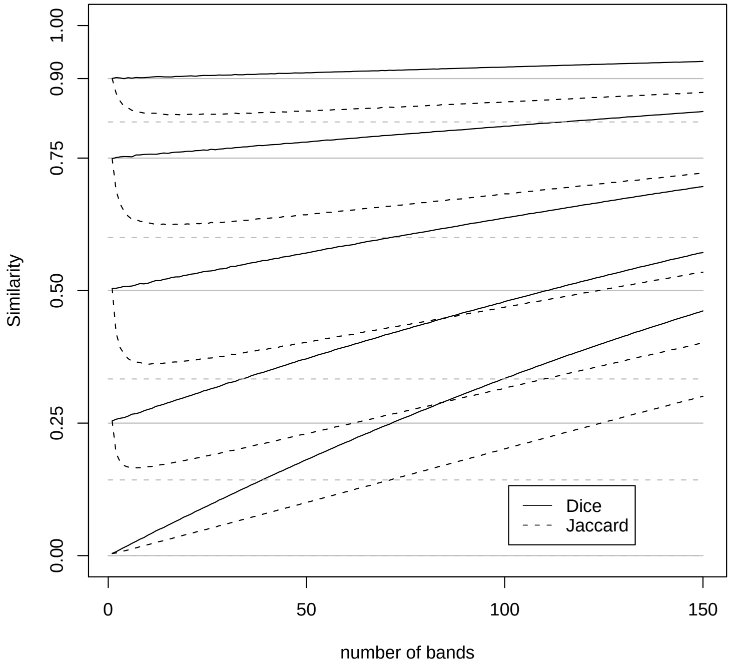


Figure 4.2: Average Dice and Jaccard coefficients as a function of number of bands for simulated AFLP profiles with genetic similarities 0, 0.25, 0.5, 0.75 and 0.9. Fragments are sampled from the fld $\boldsymbol{F}_S$ with scoring range 51-550 (see section 4.3). Horizontal lines indicate the true genetic similarity. Equal numbers of bands for the pairs of profiles are taken.

4.2.3 Collisions and band lengths

Gort, Koopman, and Stein (2006) studied the collision problem from a probabilistic point of view. They estimated the distribution of fragment lengths (fld) in different ways, based on theoretical considerations, in-silico AFLP on DNA sequence data, and empirical band data. It was found that shorter fragments occur more often than longer ones. Based on the fld, they calculated collision probabilities, and concluded that for plants like lettuce a profile with only 19 bands likely contains a collision, whereas a quarter of the bands may be collision infested for large band counts (say 100).

The present study focuses on the relationship between collision occurrence and band length. The finding of Vekemans et al. (2002) that shorter bands are more prone to collisions than longer bands, is elaborated and quantified. For application of our results we selected a subset of profiles from the lettuce study, to cover a wide range of band counts: *L. tenerrima* 9387 with 28 bands, *L. serriola* 14314 with 45 bands, *L. sativa* "Norden" with 56 bands, and *L. tatarica* $W9530$ with 100 bands. The profiles of *L. serriola* and *L. sativa* are visible in figure 4.1.
Section 4.3 contains notation and earlier results. In section 4.4 collision probabilities for single bands are calculated as a function of band length. Calculations are simplest with known fragment count in a profile (§4.4.1), but in practice only the band count (§4.4.2), and sometimes the band lengths (§4.4.3) are known, as in the lettuce study. Higher order collisions are studied in §4.4.4. In section 4.5 the total number of collisions within a profile is estimated, given the band lengths. The results are applied to arrive at modified Dice and Jaccard coefficients. Notice that in the estimation of genetic similarity both comigration of fragments between profiles and collision are important, whereas the main topic of this paper is collision. The results on collision allow us to estimate the extent of size homoplasy. Section 4.6 contains conclusions and discusses some further topics. Appendix 4.A describes some collision calculations in more detail.

4.3 Notation, assumptions, and earlier results

The AFLP technique produces profiles, containing bands at approximately discrete band positions, representing DNA fragments of specific lengths. In the following, notation for a profile is introduced:

N = total number of observable band lengths, e.g. $N = 500$ if band lengths 51-550 are observed; the range 51-550 is called the scoring range;

j = index for band length; $j = 1, \ldots, N$; the minimum observable length l_{min}, e.g. 51, is indexed as 1, the maximum l_{max}(e.g. 550) as N; a band with length index j will be referred to as a band with length j;

$\boldsymbol{y} = (y_1, \ldots, y_N)$ = vector of binary band length scores, with $y_j = 0$ if no band is present and $y_j = 1$ if a band is present with length j;

$n = \sum_{j=1}^{N} y_j$ = number of bands.

Not directly observed are the following:

$\boldsymbol{k} = (k_1, \ldots, k_N)$ with k_j = number of fragments of length j; notice that $\{y_j = 0\} \Leftrightarrow \{k_j = 0\}$, i.e. no band of length j means that no fragment of length j was amplified, and $\{y_j = 1\} \Leftrightarrow \{k_j \geq 1\}$, i.e. a band of length j means that at least one fragment of length j was amplified;

$m = \sum_{j=1}^{N} k_j$ = number of fragments in the profile;

$c = m - n$ = number of collisions.

In the first step of AFLP, restriction enzymes cut the DNA into fragments. We call this set of fragments the population of candidate fragments. The lengths of these candidate fragments form a fragment length distribution (fld). The sampling of fragments from this population takes place at the third step of the procedure, when primers select fragments for amplification and visualization. We assume that the selected fragments are a random sample from the population of candidate fragments. The sampling of fragment lengths from the fld is treated as sampling with replacement, since the sample size is small compared to the population size (see section 4.6). The visualization step of the AFLP procedure also comprises truncation of the sample, since only fragments with lengths within the scoring range appear in the profile as bands. The following notation is used:

p_j = the probability that a fragment, drawn at random from the population of candidate fragments, has length index j;

$\boldsymbol{F} = (p_1, \ldots, p_N)$ = a discrete fld;

$b_j = P(y_j = 1)$ = band probability for length j; probability that in a sample of m fragments from fld $\boldsymbol{F}$ a band with length j is observed; notice that $1 - b_j = (1 - p_j)^m$.

According to Gort et al. (2006), the fld $\boldsymbol{F}$ is monotonically decreasing, so that smaller fragments occur more often than longer fragments. When the usual restriction enzymes *Mse*I and *Eco*RI are used, the fld mainly depends on the fraction of GC nucleotides in the genome, as these enzymes cut the DNA at specific AT-rich restriction sites. Therefore, genomes with a lower GC content will be more frequently cut than GC rich genomes yielding more short fragments. Results on collisions for the following estimated fld's were compared:

$\boldsymbol{F}_S$ = fld derived from an in-silico AFLP procedure on the genome sequence of *Arabidopsis thaliana*; *A. thaliana* has a GC-content of 36%, and is thus representative for plants with GC-content ≈ 36%; ratio $p_1/p_{500} \approx 28$;

$\boldsymbol{F}_{T_1}$ = fld derived from theoretical considerations by Innan et al. (1999), assuming GC-content 50%; flatter than $\boldsymbol{F}_S$ with $p_1/p_{500} \approx 8$;

$\boldsymbol{F}_{T_2}$ = idem, assuming GC-content 36%; steeper than $\boldsymbol{F}_S$ with $p_1/p_{500} \approx 228$;

$\boldsymbol{F}_A$ = empirical estimate of the fld based on the binary scores $\boldsymbol{y}$ of the profile;

$\boldsymbol{F}_U$ = uniform fld; highly unrealistic fld, included as reference, giving lower bounds on collision probabilities (Munford, 1977).

The fld plays an important role in the estimation of collision probabilities. Therefore, careful determination of the fld is needed before collision calculations can be performed. The described estimators of the fld give reasonable options for most situations. If *a priori* information on GC-content and sequence data from a related organism with comparable GC-content are available, an estimator like $\boldsymbol{F}_S$ from an in-silico AFLP approach can be used. Without sequence data but with GC-information, an estimator like $\boldsymbol{F}_T$ can be used. The most general approach is to estimate the fld directly from the empirical AFLP data using $\boldsymbol{F}_A$, but band length information is needed.

Vector $\boldsymbol{k}$, given the fragment count m and fld $\boldsymbol{F}$, has a multinomial(m;$\boldsymbol{F}$) distribution. The probability of no collision $P_{0|m} = P(k_1 \leq 1, \ldots, k_N \leq 1 \mid m)$ is a probability from the multinomial cumulative distribution function.

The collision problem, i.e. the case that the band count n is smaller than m, is known as an occupancy problem in probability theory (Feller, 1968, p. 38). The distribution of the band count n, given m and $\boldsymbol{F}$, is a generalized occupancy distribution (Chakraborty, 1993). It may also be formulated for the number of collisions c, since $P(c = c_0|m = m_0) = P(n = m_0 - c_0|m = m_0)$. The generalized occupancy distribution is generally intractable, but can be approximated by a binomial distribution using the easily obtained first two moments. In practice not the fragment count, but the band count is observed. Therefore, the interest is in $P(m|n)$, not in $P(n|m)$. The former can be calculated by first applying Bayes' rule, and next approximating the resulting occupancy probabilities.

4.4 Probability of no collision per band

4.4.1 Probability of no collision per band, given the fragment count

Theory

The probability of no collision for a band with length j given the unobservable fragment count m is

$$P_{0|m}(j) = P(k_j = 1 \mid m;\ k_j \geq 1) = \frac{P(k_j = 1 \mid m)}{1 - P(k_j = 0 \mid m)}. \tag{4.1}$$

Since k_j given m has a Binom(m, p_j) distribution, $P_{0|m}(j)$ is the ratio of two binomial probabilities: $P_{0|m}(j) = \frac{mp_j(1-p_j)^{m-1}}{1-(1-p_j)^m}$, which can be written as

$$P_{0|m}(j) = m\frac{p_j}{1-p_j} / \frac{b_j}{1-b_j}. \tag{4.2}$$

We propose to use the ratio R_m of largest over smallest collision probability given the fragment count m as measure for the effect of band length on collision probability. Since the fld's are monotonically decreasing ($p_j > p_{j+1}$), the largest collision probability occurs for the smallest fragment. Therefore, the ratio of largest over smallest collision probability is

$$R_m = \frac{1 - P_{0|m}(1)}{1 - P_{0|m}(N)}. \tag{4.3}$$

Results

Figure 4.3 shows $P_{0|m}(j)$ as a function of band length for fragment counts $m = 10, 20, \ldots, 140$ (at $N = 500$) in subplots 1, 2, 3 and 4 for the four fld's $\boldsymbol{F}_U$,

$\boldsymbol{F}_{T_1}$, $\boldsymbol{F}_S$, and $\boldsymbol{F}_{T_2}$. Observe that, for $\boldsymbol{F}_U$, a small fragment will have a collision probability close to 7% for $m = 20$, 18% for $m = 50$, and 35% for $m = 100$. For the longest fragments these probabilities drop to approximately 0.3%, 0.7%, and 1.5% respectively. Therefore, longer bands may be considered more reliable.
In table 4.1 R_m is shown for all combinations of fld ($\boldsymbol{F}_{T_1}$, $\boldsymbol{F}_S$, $\boldsymbol{F}_{T_2}$), N (with values 400, 500, and 600), and fragment count m (with values 20, 50 and 100). Observe that the effect of the fld is large: at $N = 500$ and m=50 R_m is 7.7 for $\boldsymbol{F}_{T_1}$, 25.9 for $\boldsymbol{F}_S$, and 209.3 for $\boldsymbol{F}_{T_2}$. For $m = 20$ these values are 7.9, 27.0, and 221.8, and for $m = 100$ they are 7.4, 24.1, and 189.1. The fragment count m, although important for the absolute probabilities of collision, has only a mild influence on the probability ratio R_m. A higher number of band positions N results in higher ratios R_m, as might be expected: e.g. for $\boldsymbol{F}_S$ and $m = 50$, for $N = 400$, 500, and 600 probability ratios of 15.3, 25.9, and 43.3 are found.

			N=400			N=500			N=600		
		F_U	F_{T_1}	F_S	F_{T_2}	F_{T_1}	F_S	F_{T_2}	F_{T_1}	F_S	F_{T_2}
	$m = 20$	1.0	5.2	16.0	74.7	7.9	27.0	221.8	11.9	45.1	658.4
R_m	$m = 50$	1.0	5.1	15.3	70.5	7.7	25.9	209.3	11.7	43.3	621.4
	$m = 100$	1.0	4.9	14.2	63.7	7.4	24.1	189.1	11.3	40.3	561.4
	$n = 20$	1.0	5.2	15.8	73.9	7.9	26.8	219.2	11.9	44.8	648.3
R_n	$n = 50$	1.0	5.1	15.1	68.8	7.7	25.5	204.2	11.6	42.8	605.6
	$n = 100$	1.0	4.8	13.5	57.9	7.3	23.1	172.3	11.1	38.7	511.9

Table 4.1: Ratio of largest over smallest collision probability, given the fragment count (R_m) and given the band count (R_n)

4.4.2 Probability of no collision per band, given the band count

Theory

In practice the number of bands n in a profile is known, not the number of fragments m. Therefore, the probability of no collision for a band with length j, given n bands in total, is of interest. Since $m \geq n$, the collision probabilities will be at least as large as the values found in section 4.4.1.
The probability of no collision for a band at position j given n bands is

$$P_{0|n}(j) = P(k_j = 1 \mid n; k_j \geq 1) = \frac{P(k_j = 1 \mid n)}{1 - P(k_j = 0 \mid n)}. \tag{4.4}$$

Unlike the situation in section 4.4.1, the probabilities $P(k_j \mid n)$ in the numerator and denominator of (4.4) are not from a known distribution. They are handled in the same way: first, condition on the number of fragments, and next, use Bayes' rule to switch the order of argument and condition. In the end expressions containing probabilities from known distributions appear:

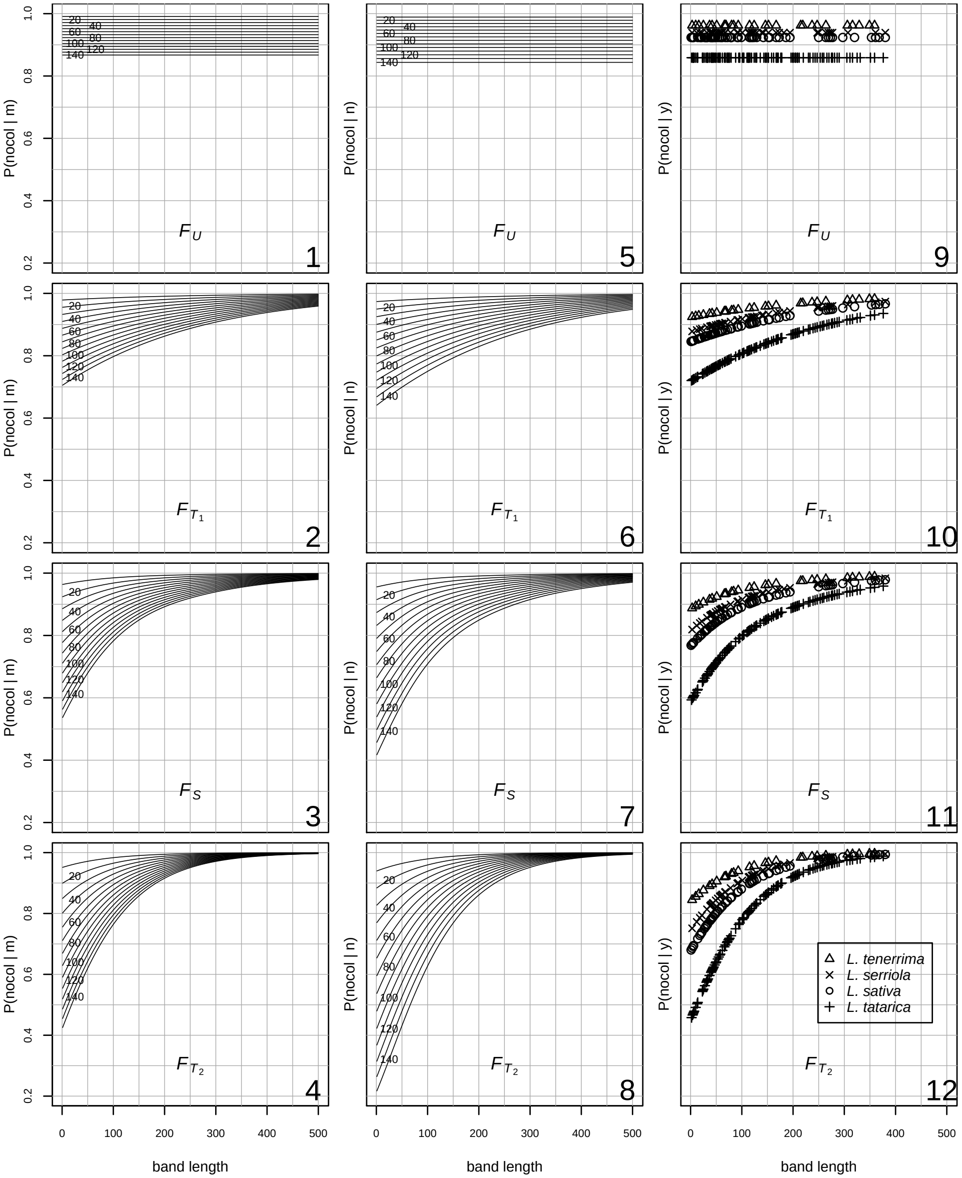


Figure 4.3: Probability of no collision as function of band length, given the fragment count m ($P_{0|m}(j)$), given the band count n ($P_{0|n}(j)$), and given the binary band scores y ($P_{i|\boldsymbol{y}}(j)$) for different fld's. In the last column AFLP examples from the lettuce study are examined. In plots 1-4 the numbers $20, 40, \ldots, 140$ are fragment counts, whereas in plots 5-8 they are band counts.

1. $Bin^m_{p_j}(k_0) = P(k_j = k_0 \mid m)$ is the binomial probability of k_0 successes out of m with success rate p_j;
2. $Occ^m_{\boldsymbol{F}}(n) = P(n \mid m)$ is the generalized occupancy probability that n cells are occupied if m balls are distributed over cells with cell probabilities from fld $\boldsymbol{F}$.

The resulting probability of no collision is

$$P_{0|n}(j) \quad = \quad \frac{\sum_{i \geq n} Occ^{i-1}_{\boldsymbol{F}^{-j}}(n-1)\ Bin^i_{p_j}(1)}{\sum_{i \geq n}\ (Occ^i_{\boldsymbol{F}}(n) - Occ^i_{\boldsymbol{F}^{-j}}(n)\ Bin^i_{p_j}(0))} \tag{4.5}$$

where $\boldsymbol{F}^{-j}$ is the rescaled fld $\boldsymbol{F}$ with the j^{th} fragment length omitted. For a detailed derivation see appendix 4.A.
Calculation of probabilities from the generalized occupancy distribution in (4.20) is troublesome. The distribution is therefore approximated by a binomial distribution with correct first two moments.
To express the effect of band length, the ratio R_n of largest over smallest collision probability given the band count is calculated as:

$$R_n = \frac{1 - P_{0|n}(1)}{1 - P_{0|n}(N)}. \tag{4.6}$$

Results

In figure 4.3 $P_{0|n}(j)$ is shown in subplots 5, 6, 7 and 8 for the four fld's, and for band counts $n = 10, 20, \ldots, 140$ with $N = 500$. The same conclusions are drawn as in section 4.4.1, but the collision probabilities are larger. For fld $\boldsymbol{F_S}$ a small band will have a collision probability close to 0.08 with $n = 20$ bands (was 0.07 for $m = 20$), 0.21 for $n = 50$ (was 0.18), and 0.42 for $n = 100$ (was 0.35). The probabilities are 0.003, 0.009, and 0.019 for the longest bands, respectively.
Table 4.1 contains the ratio of largest to smallest collision probability R_n for the four fld's, $N = 400, 500$ and 600, and band counts $n = 20, 50$ and 100. The same conclusions as in §4.4.1 are drawn. Notice that for fld $\boldsymbol{F_S}$ with $N = 500$ a collision for the shortest band is close to 25 times as likely as for the longest band. This number is only mildly influenced by the band count.

4.4.3 Probability of no collision per band, given the band lengths

Theory

In this section we suppose that the band lengths are known. This information can be used to get a more refined estimate of the collision probability for a band with a specific length. The probability of no collision for a band with length j given the binary band length vector $\boldsymbol{y}$ (with band count n) is

$$P_{0|\boldsymbol{y}}(j) = P(k_j = 1 \mid \boldsymbol{y}) = \sum_{i \geq n} P(k_j = 1 \mid \boldsymbol{y};\ m = i)\ P(m = i \mid \boldsymbol{y}), \tag{4.7}$$

by conditioning on the number of fragments. Notice that $P_{0|\boldsymbol{y}}(j)$ must be 0, if $y_j = 0$ (no band with length j).

Using Bayes' rule, $P(m = i \mid \boldsymbol{y})$ from (4.7) is written as

$$P(m = i \mid \boldsymbol{y}) = \frac{P(\boldsymbol{y} \mid m = i) \times P(m = i)}{\sum_{l \geq i} P(\boldsymbol{y} \mid m = l) \times P(m = l)}. \tag{4.8}$$

Use Bayes' rule for the probability $P(k_j = 1 \mid \boldsymbol{y};\ m = i)$ in (4.7) as well:

$$P(k_j = 1 \mid \boldsymbol{y};\ m = i) = \frac{P(\boldsymbol{y} \mid k_j = 1;\ m = i)\ P(k_j = 1 \mid m = i)}{P(\boldsymbol{y} \mid m = i)}. \tag{4.9}$$

Combining (4.8) and (4.9) gives:

$$P_{0|\boldsymbol{y}}(j) = \sum_{i \geq n} \frac{P(\boldsymbol{y} \mid k_j = 1; m = i)\ P(k_j = 1 \mid m = i)\ P(m = i)}{\sum_{l \geq i} P(\boldsymbol{y} \mid m = l)\ P(m = l)}. \tag{4.10}$$

The probabilities making up the rhs of (4.10) can now be evaluated :

1. $P(\boldsymbol{y} \mid m = l)$ is a multinomial tail probability (because $\{y_j = 0\} \Leftrightarrow \{k_j = 0\}$, and $\{y_j = 1\} \Leftrightarrow \{k_j \geq 1\}$, see section 4.3), with l fragment lengths randomly drawn from the known fld $\boldsymbol{F}$. This probability can be approximated with high accuracy with a saddlepoint approximation, as described in Butler and Sutton (1998).

2. $P(\boldsymbol{y} \mid k_j = 1;\ m = i)$ is a multinomial tail probability, but now omitting band length j from $\boldsymbol{y}$, with $i - 1$ remaining fragments distributed over $N - 1$ remaining positions. Again, the probability can be approximated with a saddlepoint approximation.

3. $P(k_j = 1 \mid m = i)$ is a binomial probability.

4. $P(m = i)$ is a prior distribution of fragment counts. We take the uniform distribution, although prior information on the number of fragments is available. In section 4.6 different sources of prior information are described and consequences for straightforward inclusion into an informative prior distribution are discussed. We conclude that straightforward inclusion of this information into a highly informative prior distribution may lead to incorrect results.

As before, the ratio R_y of largest over smallest collision probability, now given the band lengths, expresses the effect of the band length:

$$R_y = \frac{1 - P_{0|\boldsymbol{y}}(j_{min})}{1 - P_{0|\boldsymbol{y}}(j_{max})} \tag{4.11}$$

with j_{min} the smallest and j_{max} the largest observed fragment length.

Results

The four selected AFLP profiles from the lettuce study are analysed. $\boldsymbol{F}_S$ is a reasonable fld, since all four species have GC-contents close to 36%. Figure 4.3 shows the probability of no collision for each observed band, as a function of band length, for the four fld's in subplots 9, 10, 11 and 12. Notice again the large effect of fld, necessitating the determination of a proper fld. The larger the number of bands, the larger the collision probabilities, as expected. The shortest band of *L. tatarica* has a collision probability close to 0.4.

The ratio R_y for the selected AFLP profiles, using fld $\boldsymbol{F}_S$, is 10.0 for *L. tenerrima* ($j_{min} = 3, j_{max} = 359$), 10.9 for *L. serriola*(3, 380), 10.9 for *L. sativa* (1, 380) and 9.8 for *L. tatarica* (2, 376). So, the shortest fragments are approximately 10 times more likely to contain collisions than the longest fragments.

4.4.4 Higher order collisions per band

In the previous subsections the probability of no collision was calculated. If a collision does occur, it might be a single collision (2 fragments of the same length), a double collision (3 fragments of the same length), a triple collision (4 fragments of the same length), or a multiple collision (5 fragments or more of the same length). With the theory developed in the previous sections, probabilities of specific higher order collisions can be calculated as well. Again, situations for known fragment count m, known band count n, or known band lengths $\boldsymbol{y}$ are discriminated. The collision probabilities are denoted as $P_{i|m}(j)$, $P_{i|n}(j)$, and $P_{i|\boldsymbol{y}}(j)$, with i the index for the order of the collision, e.g. $P_{2|m}(j)$ is the probability of a second order collision for a band with length j, given m fragments in the profile. Results are plotted in figure 4.4 for $P_{i|m}(j)$ with $m = 20, 50, 100, 120$ (subplots 1-4), and for $P_{i|n}(j)$ with $n = 20, 50, 100, 120$ (subplots 5-8) for profiles with $N = 500$ and fld $\boldsymbol{F}_S$. The collision probability is split into probabilities for single ($i = 1$), double ($i = 2$), triple ($i = 3$), and multiple ($i > 3$) collisions. Results for $P_{i|\boldsymbol{y}}(j)$ for the AFLP examples on lettuce are given in subplots 9-12, where the vertical black lines correspond to the observed bands.

Notice from figure 4.4 that the probability of a higher order collision is negligible for long bands. However, a short band in an AFLP containing a high number of bands (e.g. $n = 120$) has a probability of a double or higher order collision close to 0.2. The shortest bands in the AFLP example on *L. tatarica* have probabilities of a single, double, and triple collision close to 0.27, 0.1, and 0.025, respectively.

4.5 Number of collisions in a profile, given the band lengths

4.5.1 Theory

Whereas in section 4.4 the collision probability for a single band in a profile was the quantity of interest, now the focus is on the total number of collisions c in a profile. In Gort et al. (2006) c was estimated, assuming known fragment count m, and known band count n. In this section the distribution of c is calculated, using the binary band information $\boldsymbol{y}$, allowing more refined estimates. The probability of interest is $P(c \mid \boldsymbol{y})$, which can also be written as $P(m \mid \boldsymbol{y})$, with $m = c + n$, and number of bands $n = \sum_{j=1}^{N} y_j$. Straightforward application of Bayes' theorem

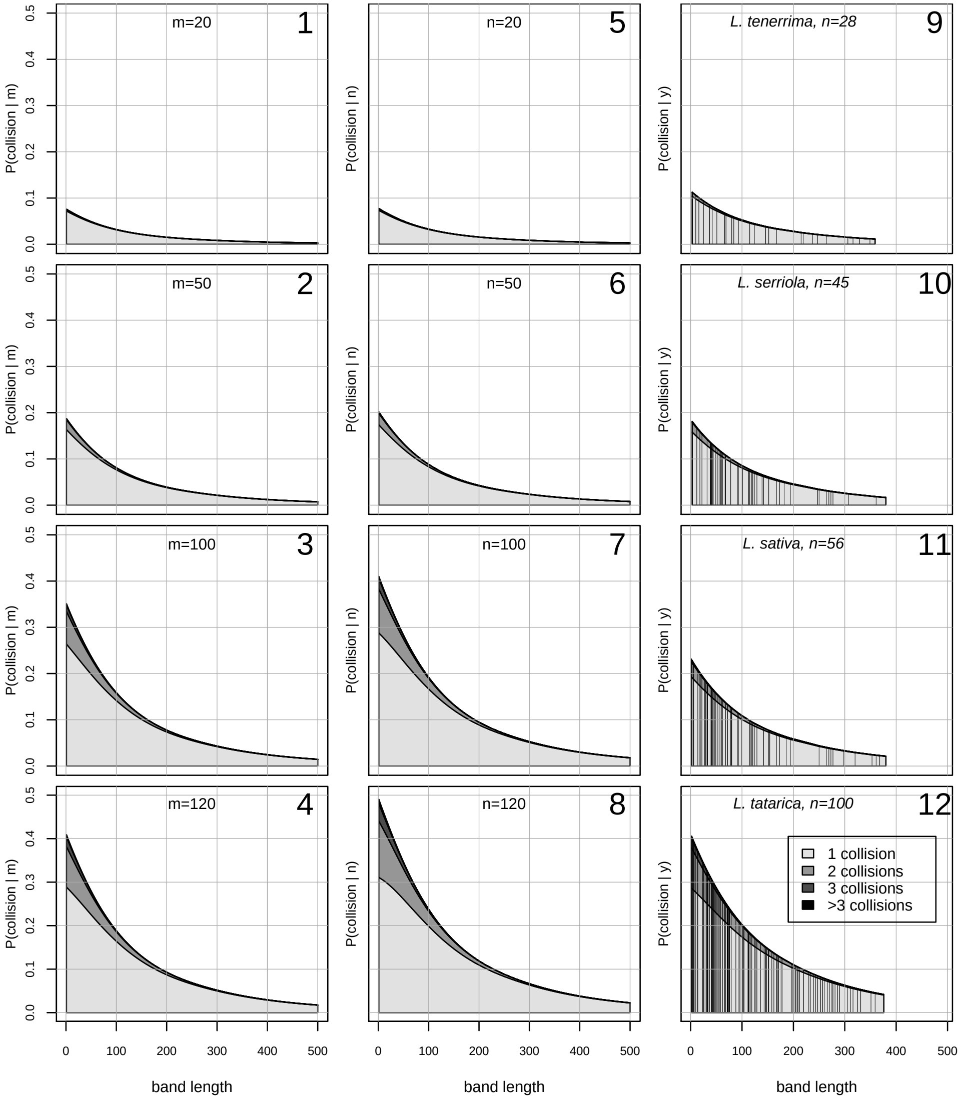


Figure 4.4: Probability of higher order collisions per band length given the fragment counts $m = 20, 50, 100, 120$ ($P_{i|m}(j)$), given the band counts $n = 20, 50, 100, 120$ ($P_{i|n}(j)$), and given the band lengths for four AFLP examples on lettuce ($P_{i|y}(j)$), based on fld $\boldsymbol{F}_S$.

gives

$$P(c \mid \boldsymbol{y}) = \frac{P(\boldsymbol{y} \mid m) \times P(m)}{\sum_{i \geq n} P(\boldsymbol{y} \mid m = i) \times P(m = i)} = \frac{P(\boldsymbol{y} \mid m)}{\sum_{i \geq n} P(\boldsymbol{y} \mid m = i)} \tag{4.12}$$

assuming, as in §4.4.3, that *a priori* all fragment counts are equally likely. Numerator and denominator of (4.12) are (sums of) multinomial (tail) probabilities. As in §4.4.3 these probabilities are approximated with a saddlepoint approximation. As a result the distribution of c is approximated, from which expectation and s.d. are easily derived.

4.5.2 Results

The profiles on lettuce are, again, analysed. Table 4.2 contains the expectation and s.d. of the number of collisions. Results for the four fld's are given, but also for the empirical fld $\boldsymbol{F}_A$, estimated from the profile itself (see 4.3). The more skewed distributions result in higher collision counts. $\boldsymbol{F}_A$ leads to collision counts which are in the range given by $\boldsymbol{F}_{T_1}$ and $\boldsymbol{F}_{T_2}$. The profile for *L. tatarica* with 100 bands is likely to contain 23.7 collisions based on $\boldsymbol{F}_S$, whereas the empirical estimate $\boldsymbol{F}_A$ results in 22.2 collisions.

Species	n	F_U	F_{T_1}	F_S	F_{T_2}	F_A
L. tenerrima	28	1.09 (1.07)	1.33 (1.19)	1.48 (1.26)	1.69 (1.36)	1.32 (1.18)
L. serriola	45	2.86 (1.76)	3.81 (2.06)	4.33 (2.22)	5.06 (2.44)	4.29 (2.21)
L. sativa	56	4.51 (2.24)	6.31 (2.71)	7.60 (3.02)	9.41 (3.45)	7.85 (3.08)
L. tatarica	100	15.62 (4.36)	20.51 (5.17)	23.70 (5.70)	28.49 (6.43)	22.19 (5.44)

Table 4.2: Estimated numbers of collisions (s.d.) for examples of AFLP profiles from *Lactuca*

4.5.3 Modified Dice and Jaccard coefficients

The theory of section 4.5.1 can be used to arrive at modified Dice and Jaccard coefficients for estimation of genetic similarity from AFLP profiles. Given are the profiles from two genotypes with band counts n_1 $(= a + b)$ and n_2 $(= a + c$, with a, b and c defined in §4.2.2). Let n_{1+2} $(= a + b + c)$ be the band count, by combining the two profiles into a single profile; $n_{12} = n_1 + n_2 - n_{1+2}$ $(= a)$ is the number of shared bands. Then the Dice coefficient $D = 2n_{12}/(n_1 + n_2)$, and Jaccard coefficient $J = n_{12}/n_{1+2}$.

We propose the modified Dice and Jaccard coefficients $D_{mod} = 2m_{12}/(m_1 + m_2)$ and $J_{mod} = m_{12}/m_{1+2}$, with m_1 and m_2 the estimated fragment counts for the two profiles, m_{1+2} the estimated fragment count for the combination of the two profiles, and $m_{12} = m_1 + m_2 - m_{1+2}$.

Since the calculations using the saddlepoint approximation are very slow, a second method for estimation of m is proposed, based on the EM-algorithm (Dempster, Laird, & Rubin, 1977). The EM-algorithm, however, does not give the precision of the estimate. For the EM-algorithm we treat the problem as an incomplete data problem, where the unobserved fragment counts k_j are missing. The band count n is taken as a starting value of m. For the E-step we calculate the expected conditional log-likelihood $ECLL$ of k_j, given y_j. Assuming a Poisson distribution for k_j with mean mp_j, we arrive at $ECLL = \sum_j (\kappa_j log(mp_j) - mp_j)$, leaving out terms not depending on m. In this expression

$$\kappa_j = E(k_j|y_j) = \begin{cases} 0 & \text{if} \quad y_j = 0 \\ mp_j/(1 - e^{mp_j}) & \text{if} \quad y_j = 1 \end{cases}$$

are the pseudodata, updated by filling in the current value of m. The $ECLL$ can be recognized as a Poisson log-likelihood for κ_j (besides terms not depending on m), which in the M-step can be maximised by fitting a generalized linear model for κ_j with Poisson distribution, log-link and systematic part of the model $log(m) + log(p_j)$ with offset $log(p_j)$ (McCullagh & Nelder, 1989). From the m.l. estimated intercept $log(m)$, an updated estimate of m is obtained by exponentiation.

A small simulation study (genetic similarity $S = 0.5$; $n = 10, 60, 120$; 2000 replicates), using the EM-algorithm for estimation of the expected fragment counts, showed that the estimators behave nicely: for $n = 10$ the average $D_{mod} = 0.501$ vs $D = 0.516$, for $n = 60$ $D_{mod} = 0.497$ vs $D = 0.583$, and for $n = 120$ $D_{mod} = 0.498$ vs $D = 0.661$, so the bias seems to be removed completely.

For the selected species *L. tenerrima*, *L. serriola*, *L. sativa* and *L. tatarica*, we find the following matrices of pairwise Dice coefficients M_D and modified Dice coefficients $M_{D_{mod}}$:

$$M_D = \begin{pmatrix} . & 0.25 & 0.19 & 0.19 \\ 0.25 & . & 0.69 & 0.28 \\ 0.19 & 0.69 & . & 0.31 \\ 0.19 & 0.28 & 0.31 & . \end{pmatrix}, \quad M_{D_{mod}} = \begin{pmatrix} . & 0.15 & 0.061 & 0.049 \\ 0.15 & . & 0.64 & 0.086 \\ 0.061 & 0.64 & . & 0.090 \\ 0.049 & 0.086 & 0.090 & . \end{pmatrix}.$$

The estimates of genetic similarity based on D_{mod} are noticeably smaller than those based on D, as expected. Stronger corrections are found for profiles with more bands, e.g. the similarity for *L. ser.* and *L. tat.* $D_{24} = 0.28$ is corrected into 0.086, whereas for *L. ten.* and *L. ser.* $D_{12} = 0.25$ changes to 0.15. We see stronger corrections for profiles with less genetic similarity, e.g. for *L. sat.* and *L. tat.* $D_{34} = 0.31$ is modified into 0.090, whereas for *L. ser.* and *L. sat.* $D_{23} = 0.69$ becomes 0.64. Notice that the order of modified similarities may differ from the original order, e.g. D_{12} is the fourth largest, but $D_{mod,12}$ is second largest. A paper on modified similarity coefficients is in preparation.

4.6 Conclusions and discussion

In this paper we studied the relationship between collision probability and band length in AFLP. The reason that band lengths relate to collision probabilities is the non-uniformity of the fld. Short fragments are more frequent, and, hence, short bands are more likely to contain collisions. Collision probabilities for individual band positions were calculated. Probabilities of higher order collisions were calculated as well. The total number of fragments or bands has a strong impact on the absolute values of the collision probabilities. The ratio of largest

to smallest collision probability over the band positions is mainly governed by the fld, with higher ratios for more skewed distributions. The expected collision count in a profile with standard deviation was estimated for a number of examples.
The collision problem has an impact on all uses of AFLP in which bands are assumed to represent single fragments of DNA. As an example we examined simple measures for genetic similarity. A small simulation study suggested that the proposed modified Dice and Jaccard coefficients estimate the genetic similarity unbiasedly, but more work is needed to study other properties of these estimators. Based on our present findings a number of suggestions can be made for the practical use of AFLP.

- It is strongly suggested that in AFLP analyses band lengths are reported as much as possible, since band lengths contain valuable information about collision probabilities.
- Researchers should score all bands within a profile, or at least mention how many bands are present. If bands are ignored (e.g. monomorphic bands, occurring in all genotypes of the gel) a proper judgement of the extent of the collision problem is not possible.
- For procedures such as linkage map construction, where the bands are assumed to represent single fragments, a safe procedure is to use highly selective primers, aiming at not more than 20 bands per profile. If more bands per profile are produced, knowledge of the band length allows the researcher to pinpoint possibly problematic bands.
- Researchers tend to construct profiles with many bands, up to 100 or more. They should realize that a quarter or more of those bands may contain collisions, crowded at the shorter band positions.

In the collision calculations given the band count and given the band positions, a prior distribution of the fragment count is needed. In the calculations we take a rather naive approach by assuming a uniform prior distribution. Usually, however, *a priori* information is available. Some sources of information are related to the AFLP procedure:

- Restriction enzymes. Two restriction enzymes, usually *Mse*I and *Eco*RI, are used to cut the total DNA into fragments. These fragments make up the population of candidate fragments. Different restriction enzymes will lead to different and differently sized populations of candidate fragments.
- Primers. The selective nucleotides of the primers determine the selection of fragments to be amplified. Each additional nucleotide roughly yields a fourfold reduction in the number of fragments. One of the primers is labeled, and only fragments including this primer will be visualized.
- Scoring range. Only bands with lengths in the scoring range are scored. Smaller scoring ranges lead to less fragments.
- Flaws in DNA amplification. Due to competition in the polymerase chain reaction during the amplification procedure not all selected fragments may be

properly amplified. As a result some fragments may remain undetected. This problem is more serious for larger genomes and for larger fragments.

Other sources of information are related to the genome:

- Genome size. Larger genomes yield larger populations of candidate fragments, and therefore higher fragment counts.

- Repetitive DNA. If a selected fragment is located within a section of repetitive DNA, multiple copies of the fragment will occur. Since larger genomes generally contain more repetitive DNA, it is likely that they also yield larger proportions of fragments from repetitive stretches. Hence, the number of fragments with different nucleotide sequences may be lower than projected from the genome size.

- Zygosity. Self-pollinators like *L. sativa* are mainly homozygous, but cross-pollinators like *L. tatarica* are heterozygous for a large number of loci. Higher heterozygosity leads to a higher number of different candidate fragments.

Given the above sources of information, the question arises whether it is possible to use this information to arrive at an informative prior distribution for the number of fragments. We take *L. sativa* as an example. It has a genome size of 6 picogram of DNA (Koopman, 2002), which is about 6×10^9 basepairs (bp; 1 picogram = 0.98×10^9 bp, Bennett, Bhandol, and Leitch (2000)). Since it is homozygous, the relevant amount of DNA is 3×10^9 bp. In the AFLP procedure the restriction enzymes *Mse*I and *Eco*RI were used. From an in-silico AFLP procedure on *A. thaliana* (Koopman & Gort, 2004, see), we find an average fragment length of 119 bp for all restricted fragments, leading to almost 25×10^6 candidate fragments. The fragments are of three types: Mse-Mse, Mse-Eco and Eco-Eco. Selection takes place at different steps: the primers select only 0.02% of the Mse-Eco fragments and 0.034% of the Eco-Eco fragments, only the Mse-Eco (6.9%) and Eco-Eco (0.24%) fragments are labelled, and about 50% of fragments fall within the scoring range. This results in approximately 175 fragments. Notice that the observed number of bands for *L. sativa* was 56 (table 4.2), with an estimated number of fragments equal to $56 + 7.6 = 63.6$, based on fld $\boldsymbol{F}_S$ and a uniform prior.
To assess the effect of inclusion of a-priori information on the collision count we performed a small case study, using the AFLP example on *L. sativa.* If we take as prior distribution a (discretized) normal distribution with $\mu = 175$ and $\sigma = 50$, reflecting roughly the information described above with the high σ indicating high uncertainty, the collision count becomes 7.9. A highly informative prior distribution N(175,10) results in 15.3. A prior distribution N(80,15), with mean closer to the found band count, leads to collision count of 8.2. We conclude that inclusion of a highly informative prior distribution can have a noticeable effect on the estimated collision count. Careless application of genome and AFLP procedure information as described above may lead to erroneous conclusions. Use of the uniform prior distribution should be adequate for most cases.
Our definition of a collision is the comigration of two or more fragments, originating from different loci and with different nucleotide compositions. If a diploid genotype is homozygous for an AFLP fragment, two identical DNA fragments will

be amplified, and the two fragments will be comigrating. However, we do not call this a collision. The result would be a band with an intensity that is on average higher than for a heterozygous genotype. This property is employed in codominant scoring of AFLPs. In codominant scoring the zygosity of fragments is inferred from the intensity of the band score (Piepho & Koch, 2000). Collision calculation and codominant scoring are intertwined if band intensities are studied. A paper on this topic is in preparation. The same type of problem occurs if a fragment is sampled from a repetitive stretch of DNA, so that there are multiple identical copies originating from different loci. We again do not call this collision.
Software in R (R Development Core Team, 2005) for the calculation of the collision probabilities is available from the authors.

4.7 Acknowledgements

We are grateful to Theo van Hintum from the Centre of Genetic Resources in Wageningen for his suggestion of modified similarity coefficients, and to a reviewer, whose comments helped to improve the paper.

4.A Appendix: Probability of no collision given the band count

The required probability is $P_{0|n}(j) = P(k_j = 1 \mid n = n_0,\ k_j \geq 1)$. For clarity, we write $P(k_j = 1 \mid n = n_0,\ k_j \geq 1)$ instead of $P(k_j = 1 \mid n,\ k_j \geq 1)$ as we did in equation 4.4. Work out the conditional probability:

$$\begin{aligned} P_{0|n}(j) &= P(k_j = 1,\ k_j \geq 1 \mid n = n_0) \ / \ P(k_j \geq 1 \mid n = n_0) \\ &= P(k_j = 1 \mid n = n_0) \ / \ (1 - P(k_j = 0 \mid n = n_0)) \\ &= P_1 \ / \ (1 - P_0). \end{aligned} \tag{4.13}$$

The probabilities P_1 and P_0 of (4.13) are dealt with in the same way by conditioning on the number of fragments:

$$\begin{aligned} P_1 &= \sum_{i \geq n_0} P(k_j = 1\ , m = i \mid n = n_0) \\ &= \sum_{i \geq n_0} P(k_j = 1 \mid n = n_0,\ m = i)\ P(m = i \mid n = n_0) \end{aligned} \tag{4.14}$$

Now work out the last two probabilities in (4.14).
The probability $P(m = i \mid n = n_0)$ of (4.14) is the probability of the fragment count given the band count. In Gort et al. (2006) this probability is approximated, firstly by application of Bayes' rule arriving at expression

$$P(m = i \mid n = n_0) = \frac{P(n = n_0 \mid m = i)\ P(m = i)}{\sum_{l \geq n_0} P(n = n_0 \mid m = l)\ P(m = l)} \tag{4.15}$$

As described in section 4.5 $P(n = n_0 \mid m = i)$ stems from a generalized occupancy distribution (Chakraborty, 1993), which we approximate by a binomial distribution with correct first two moments. Since we use the uniform distribution as prior distribution of m, the probabilities for the fragment lengths cancel out.
Apply Bayes rule to the first probability of (4.14) as well:

$$P(k_j = 1 \mid n = n_0, m = i) = \frac{P(n = n_0 \mid k_j = 1, m = i)\ P(k_j = 1 \mid m = i)}{P(n = n_0 \mid m = i)} \tag{4.16}$$

for which the components can be written as:

1. $P(n = n_0 \mid k_j = 1,\ m = i) = P(n' = n_0 - 1 \mid m = i - 1)$ because given i fragments and only one fragment of size j, there remain $i - 1$ fragments to be distributed of $N - 1$ positions (leaving out position j) to arrive at $n_0 - 1$ bands; this is a probability from a generalized occupancy distribution;

2. $P(k_j = 1 \mid m = i) = \binom{i}{1} p_j (1 - p_j)^{i-1}$ is an ordinary binomial probability;

3. $P(n = n_0 \mid m = i)$ is a probability from a generalized occupancy distribution.

Because the term $P(n = n_0 \mid m = i)$ vanishes and we again assume that *a priori* all fragment counts are equally likely, the result for the numerator of (4.13) is:

$$P(k_j = 1 \mid n = n_0) = \frac{\sum_{i \geq n_0} P(n' = n_0 - 1 \mid m = i - 1)\ P(k_j = 1 \mid m = i)}{\sum_{l \geq n_0} P(n = n_0 \mid m = l)} \tag{4.17}$$

The denominator of (4.13) is handled in the same way as the numerator:

$$\begin{aligned}1 - P(k_j = 0 \mid n = n_0) = 1 - \textstyle\sum_{i \geq n_0} P(k_j = 0,\ m = i \mid n = n_0) = \\ = 1 - \textstyle\sum_{i \geq n_0} P(k_j = 0 \mid n = n_0,\ m = i)\ P(m = i \mid n = n_0) \qquad (4.18)\end{aligned}$$

where only $P(k_j = 0 \mid n = n_0,\ m = i)$ is slightly different compared to $P(k_j = 1 \mid n = n_0,\ m = i)$:

$$P(k_j = 0 \mid n = n_0,\ m = i) = \frac{P(n = n_0 \mid k_j = 0,\ m = i)\ P(k_j = 0 \mid m = i)}{P(n = n_0 \mid m = i)} \qquad (4.19)$$

Now the components are:

1. $P(n = n_0 \mid k_j = 0,\ m = i) = P(n' = n_0 \mid m = i)$ again leaving out the $j - th$ category, but now distributing all i fragments over $N - 1$ remaining positions to arrive at n_0 bands; this is a probability from a generalized occupancy distribution;

2. $P(k_j = 0 \mid m = i)$ again is a binomial probability;

3. $P(n = n_0 \mid m = i)$ is the same generalized occupancy probability as before.

Piecing all elements together results in:

$$P_{0|n}(j) \quad = \quad \frac{\sum_{i \geq n_0} Occ^{i-1}_{\boldsymbol{F}^{-j}}(n_0 - 1)\ Bin^i_{p_j}(1)}{\sum_{i \geq n_0} (Occ^i_{\boldsymbol{F}}(n_0) - Occ^i_{\boldsymbol{F}^{-j}}(n_0)\ Bin^i_{p_j}(0))} \qquad (4.20)$$

where

$\boldsymbol{F}^{-j}$ is the rescaled fld $\boldsymbol{F}$ with the j^{th} fragment length omitted,

$Bin^i_{p_j}(k_0) = P(k_j = k_0 \mid i)$ is the binomial probability of k_0 successes out of i with success rate p_j,

$Occ^i_{\boldsymbol{F}}(n_0) = P(n = n_0 \mid i)$ is the generalized occupancy probability that n_0 cells are occupied if i balls are distributed over cells with cell probabilities from fld $\boldsymbol{F}$.

Chapter 5

Homoplasy corrected estimation of genetic similarity from AFLP bands, and the effect of the number of bands on the precision of estimation [1]

by Gerrit Gort, Theo van Hintum and Fred A. van Eeuwijk

5.1 Summary

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 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.

[1] Published as: Gerrit Gort, Theo van Hintum and Fred van Eeuwijk (2009). Homoplasy corrected estimation of genetic similarity from AFLP bands, and the effect of the number of bands on the precision of estimation. *Theoretical and Applied Genetics* 119:397-416

5.2 Introduction

AFLP is a DNA fingerprinting technique, that has been employed in many studies on plants (e.g. Tams, Melchinger, & Bauer, 2005), but also in studies on fungi (e.g. Mebrate, Dehne, Pillen, & Oerke, 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, Menge, Temu, Chen, & Yan, 2006), diversity studies (e.g. Berloo, Zhu, et al., 2008), and optimization of genebank management (e.g. J. Jansen & 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 errors (e.g. parts of Meudt & Clarke, 2007, and papers cited therein), but here we focus on homoplasy. Estimation of genetic similarity is biased due to size homoplasy, see figure 5.1 (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 & Clarke, 2007; Robinson & Harris, 1999). Caballero, Quesada, and Rolán-Alvarez (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 (2000b), 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, Koopman, Stein, and van Eeuwijk (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 Material 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 figure 5.1.
- 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 the Results section. Conclusions are compiled and discussed in the Discussion section. The paper ends with appendices on bootstrapping and an overview of all symbols used.

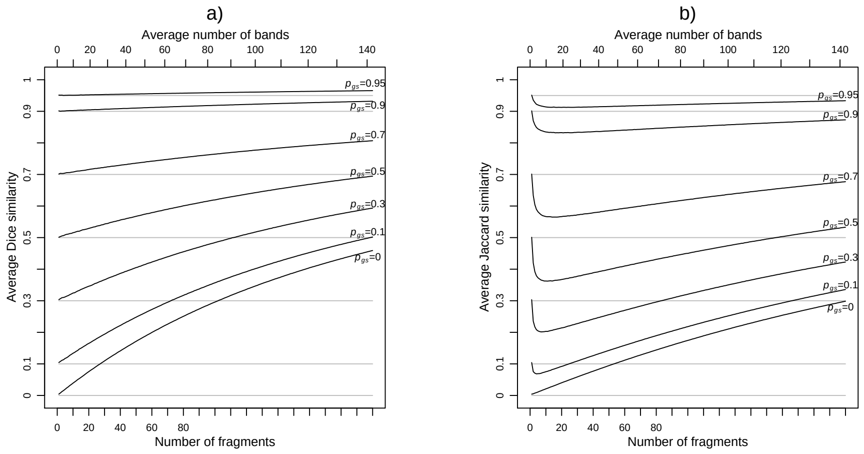


Figure 5.1: a) Average Dice, and b) average Jaccard similarities as a function of number of fragments for 100,000 simulated pairs of profiles with genetic similarities $p_{gs} = 0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95$. Fragments are sampled from $fld\ F_S$ with scoring range 51 − 500. The top axes show the average number of bands on a non-linear scale.

5.3 Material 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 *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 sixteenfold 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. 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 ($n \leq k$).

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:

$$p_{gs} = \frac{M_1}{M_1 + M_2} F_1 + \frac{M_2}{M_1 + M_2} F_2 = w_1 F_1 + w_2 F_2 \tag{5.1}$$

Notice that p_{gs} can be written as $2M_a/(2M_a + M_b + M_c)$.
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 samples 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 present in the first profile, but absent in the second $= \sum_{i=1}^{N} y_{i1}(1 - y_{i2})$;
$c =$ number of bands present in the second profile, but absent in the first $= \sum_{i=1}^{N} (y_{i1} - 1) y_{i2}$;
$d =$ number of empty positions in both profiles $= \sum_{i=1}^{N} (y_{i1} - 1)(y_{i2} - 1)$.
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, Melchinger, and Frisch (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 & 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: $J = D/(2 - D)$. 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, Katsiotis, Leggett, Loukas, and Tsakas (2004), and Tams et al. (2005).
The simple matching coefficient (Sneath & Sokal, 1973) 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, Forbes, and Smith (2007) 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 $m_1 = m_2 (= 1, \ldots, 200)$. The maximum fragment count $m = 200$ corresponds to 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 $1, \ldots, 450$ 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 $1, \ldots, 450$, 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. 5.1. 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 & Nelder, 1989) 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

$$log(-log(1 - P_i)) = log(m) + log(-log(1 - p_i)), \tag{5.2}$$

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 $p_{gs} = \frac{M_1}{M_1+M_2}F_1 + \frac{M_2}{M_1+M_2}F_2 = w_1F_1 + w_2F_2$, as in eq. 5.1. 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

$$D^{mod} = \frac{\hat{m}_1}{\hat{m}_1 + \hat{m}_2}\frac{\hat{m}_a}{\hat{m}_1} + \frac{\hat{m}_2}{\hat{m}_1 + \hat{m}_2}\frac{\hat{m}_a}{\hat{m}_2} = \frac{2\hat{m}_a}{2\hat{m}_a + \hat{m}_b + \hat{m}_c}, \tag{5.3}$$

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

$$J^{mod} = \frac{\hat{m}_a}{\hat{m}_a + \hat{m}_b + \hat{m}_c}. \tag{5.4}$$

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 $p_{gs} \geq 0$. 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

$$D^{mod}_{\bar{L}} = \frac{2\hat{m}_{\bar{L}a}}{2\hat{m}_{\bar{L}a} + \hat{m}_{\bar{L}b} + \hat{m}_{\bar{L}c}}, \quad \text{and} \quad J^{mod}_{\bar{L}} = \frac{\hat{m}_{\bar{L}a}}{\hat{m}_{\bar{L}a} + \hat{m}_{\bar{L}b} + \hat{m}_{\bar{L}c}}. \tag{5.5}$$

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

$$D_L^{mod} = \frac{2\hat{m}_{La}}{2\hat{m}_{La} + \hat{m}_{Lb} + \hat{m}_{Lc}}, \quad \text{and} \quad J_L^{mod} = \frac{\hat{m}_{La}}{\hat{m}_{La} + \hat{m}_{Lb} + \hat{m}_{Lc}}. \tag{5.6}$$

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:
$(0, 0)$: no fragment of length i at all;
$(0, 1)$: 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;
$(1, 0)$: at least one fragment from Π_b, and no fragment from Π_c and Π_a;
$(1, 1)$: 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.
For the i-th pair the likelihoods of these 4 outcomes are:
$(0, 0)$: $\ell_i = (1 - p_i)^{m_b + m_a + m_c}$
$(0, 1)$: $\ell_i = (1 - p_i)^{m_b + m_a}(1 - (1 - p_i)^{m_c})$
$(1, 0)$: $\ell_i = (1 - (1 - p_i)^{m_b})(1 - p_i)^{m_a + m_c}$
$(1, 1)$: $\ell_i = (1 - (1 - p_i)^{m_a}) + (1 - (1 - p_i)^{m_b})(1 - p_i)^{m_a}(1 - (1 - p_i)^{m_c})$.
Next, the log-likelihood of the data $LL = \sum_{i=1}^{N} log(\ell_i)$ 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

$$D_1^{mle} = \frac{2\hat{m}_a}{2\hat{m}_a + \hat{m}_b + \hat{m}_c} = \hat{w}_1\hat{p}_1 + \hat{w}_2\hat{p}_2, \tag{5.7}$$

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 $p_{gs} = \frac{2M_a}{2M_a + M_b + M_c}$, it follows $\frac{p_{gs}}{1 - p_{gs}} = \frac{M_a}{(M_b + M_c)/2}$, since we assume expected fragment counts proportional to population counts. Now, we replace m_a in the likelihood above by $\frac{p_{gs}}{1 - p_{gs}}(m_b + m_c)/2$. 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 $\frac{1}{2}(m_a + m_b)exp(l_{gs})$, with $l_{gs} = logit(p_{gs})$, yielding an estimator on the logit-scale, to be back-transformed to $D_3^{mle} = logit^{-1}(\hat{l}_{gs}) = exp(\hat{l}_{gs})/(1 + exp(\hat{l}_{gs}))$. This estimator may have better distributional properties for p_{gs} close to 0 or 1.

Precision of the estimators
The precisions of estimators $D_{\bar{L}}^{mod}$ and D_L^{mod} are determined by bootstrapping (Efron & Tibshirani, 1993), whereas for D^{mle} the precision follows from standard likelihood theory. For estimator $D_{\bar{L}}^{mod}$ 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 $(y_{11}, y_{12}), (y_{21}, y_{22}), \ldots$ $\ldots, (y_{N1}, y_{N2})$, 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 $D_{\bar{L}}^{mod}$ and D_L^{mod} we took 1000 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 & 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 $D_2^{mle} \pm se(D_2^{mle}) \times z_{1-\alpha/2}$, with $z_{1-\alpha/2}$ the $1-\alpha/2$ quantile from the standard normal distribution. For D_3^{mle} we back-transform the Wald-confidence intervals $\hat{l}_{gs} \pm se(\hat{l}_{gs}) z_{1-\alpha/2}$ using $logit^{-1}$. Besides Wald-type confidence intervals we calculated profile likelihood confidence intervals for p_{gs} (see e.g. Venzon & 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 $m_a = (m_1 + m_2)p_{gs}$, 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 ${}_c$ 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 $D^{mod}_{\bar{L}}$, D^{mod}_{L} (with 1000 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^{mle}_2 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^{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 5 selected species. We selected the 5 species to have a wide range of band counts: mean counts (± s.d.) 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, 2002) versus 36% for *A. thaliana*. The relationships between the species are visualized with UPGMA dendrograms, using dissimilarities $1 - D$, $1 - J$, and $1 - D^{mle}$.

5.4 Results

General results from the simulation study
Table 5.1 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 3 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 $m = 40$.

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

Table 5.1: 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. *Fld* F_S from *A. thaliana* is used, with $N = 450$ band positions.

Results from the simulation study for modified Dice coefficients
Table 5.2 shows the results from the simulation study for the modified Dice coef-

ficient $D_{\bar{L}}^{mod}$, using profiles without band length information. In Table 5.3 results for D_L^{mod} are given. We notice the following.

1. Almost all of the bias of the original Dice coefficient is removed. $D_{\bar{L}}^{mod}$ and D_L^{mod} slightly underestimate p_{gs} (mean observed biases -0.0018 and -0.0015, averaged over all settings of p_{gs} and m, for $D_{\bar{L}}^{mod}$ and D_L^{mod}, resp.), with largest observed bias equal to -0.0030 occurring for $D_{\bar{L}}^{mod}$ in case $p_{gs} = 0$ 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 $p_{gs} = 0.90$ 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 $p_{gs} = 0.95$ 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 $p_{gs} = 0.95$, and too narrow for estimates larger than 0.95.
3. The bootstrap standard errors of $D_{\bar{L}}^{mod}$ and D_L^{mod} are smaller for larger number of expected fragments, with the exception of $p_{gs} = 0$ and $p_{gs} = 0.1$. Hence, in the examples for $p_{gs} > 0.1$ larger fragment counts result in more precise estimates. The same can be said for the lengths of the 95% confidence intervals. If $p_{gs} = 0.1$ the smallest standard error is observed for $m = 70$.
4. The estimates $D_{\bar{L}}^{mod}$ and D_L^{mod} may become negative for small values of p_{gs}. In the table this can be seen for $p_{gs} = 0$, resulting in a negative average of D^{mod}, but it also occurs for $p_{gs} = 0.1$. For $p_{gs} = 0.3$ 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 5.2 and 5.3, showing results for the truncated versions of $D_{\bar{L}}^{mod}$ and D_L^{mod} for $p_{gs} = 0.0, 0.1,$ and 0.3. Since the truncation causes more distributional asymmetry we give medians instead of averages of $D_{\bar{L}}^{mod}$ and D_L^{mod}. For $D_{\bar{L}}^{mod}$ the bias-correction decreases the bias, but this is not always the case for D_L^{mod}. For $p_{gs} = 0$ we give the non-coverage of the (97.5%) confidence interval only at the right of $p_{gs} = 0$. For $p_{gs} = 0$ 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 $D_{\bar{L}}^{mod}$ 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 s.e.($D_{\bar{L}}^{mod}$) le s.e.(D_L^{mod}), but again differences are small. The coverage of the 95% confidence interval of $D_{\bar{L}}^{mod}$ is slightly better than that of D_L^{mod}: average absolute deviation from the nominal 5% is 0.33% for $D_{\bar{L}}^{mod}$ 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 $D_{\bar{L}}^{mod}$ has slightly better characteristics than D_L^{mod}.

Parameter			Part I Results for D_L^{mod}				
settings			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	-0.0016	-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
	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
	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

Table 5.2: 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, 1000 bootstrap resamples, *fld* F_S from *A. thaliana* with $N = 450$. 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.

Parameter			Part I Results for D_L^{mod}				
settings			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
	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
	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

Table 5.3: 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, 1000 bootstrap resamples, $fld\ F_S$ from *A. thaliana* with $N = 450$. 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.

Results from the simulation study for maximum likelihood estimators D^{mle}
Table 5.4 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 $p_{gs} \geq 0.9$ 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 $p_{gs} > 0.1$, a negligible negative bias of D_2^{mle} remains: the mean bias is 0.0015. For $p_{gs} \leq 0.1$ a small positive bias is observed, because of the necessarily non-negative value of the estimators. For $p_{gs} = 0$ the medians (not shown) are 0, and for $p_{gs} = 0.1$ 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 $0.3 \leq p_{gs} \leq 0.7$ 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 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 ∞, 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 3 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}: $p_{gs} < 0.7$ irrespective of m, and $p_{gs} \geq 0.7$ 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 $p_{gs} = 0.5$ and 0.7), making them the least

Parameter			Results for D^{mle}							
settings			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)	1.1401	.	.
	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
	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
	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

Table 5.4: 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, $fld\ F_S$ from *A. thaliana* with $N = 450$. Shown are the mean, mean standard error, and properties of 3 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. For the 4 parameter settings labeled with $^{*)}$ identical pairs of profiles were sampled (10, 348, 53, and 10 times resp.); in these cases $D_2^{mle} = 1$ with standard error 0, and we took $logit(p_{gs}) = 16$ with standard error 0.

attractive in this range.

8. For all cases with $p_{gs} \geq 0.3$, D_2^{mle} has smaller standard errors than $D_{\bar{L}}^{mod}$ 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 $D_{\bar{L}}^{mod}$ and D_L^{mod}. These results suggest that D_2^{mle} is to be preferred over the modified coefficients $D_{\bar{L}}^{mod}$ 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 $se = \sqrt{bias^2 + var}$. Using the data from the simulation study, we estimated $bias(D)$, and $var(D)$ by bootstrapping, and compared $se(D)$ with $se(D_L^{mod})$. For most cases we find $se(D) > se(D_L^{mod})$, with the most extreme outcome for $p_{gs} = 0.0$ and $m = 120$, where $se(D)$ is $4.5 \times se(D_L^{mod})$. For large values of p_{gs} ($p_{gs} = 0.95$, all m; $p_{gs} = 0.9$, $m = 40, 70$; $p_{gs} = 0.7$, $m = 40$), we find that $se(D) < se(D_L^{mod})$, but $se(D)$ is never smaller than $0.95 \times se(D_L^{mod})$. 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 5.2 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 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.

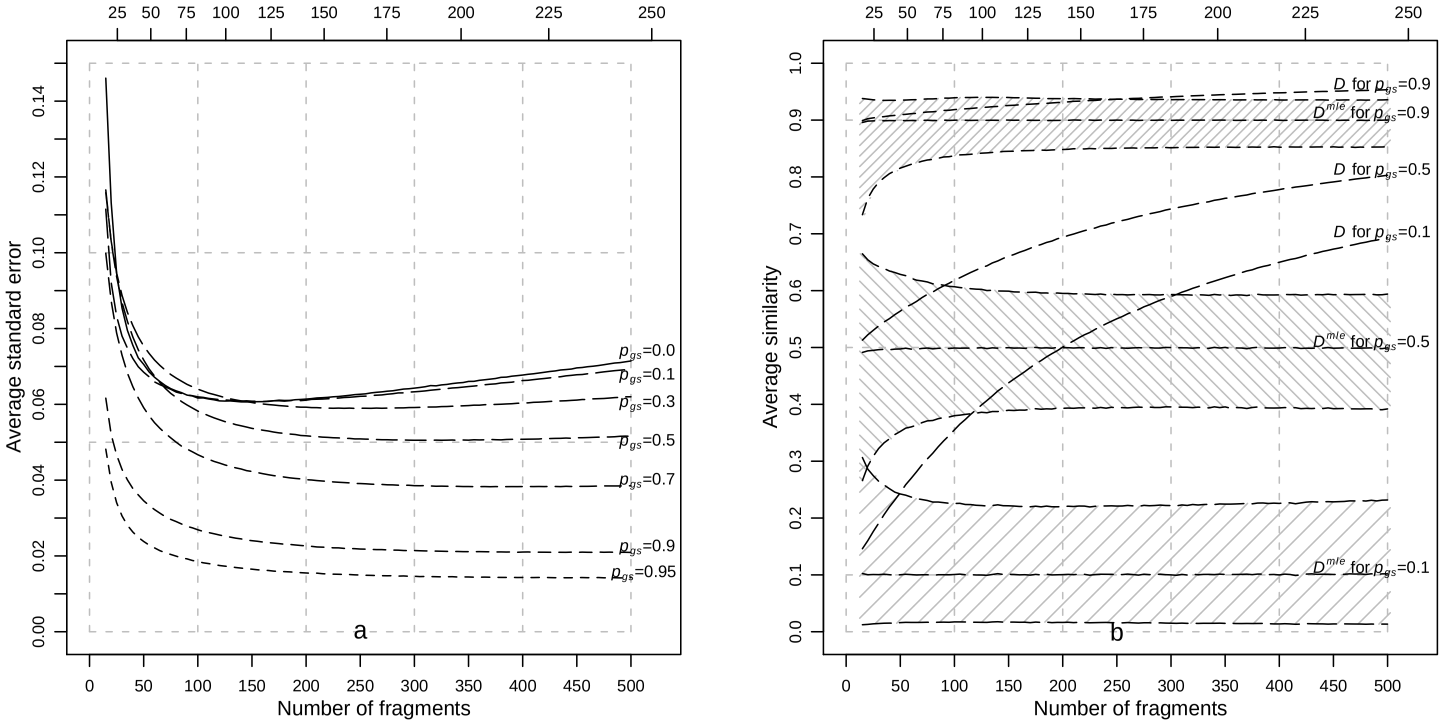


Figure 5.2: a) Average standard error of D_2^{mle}, and b) average D_2^{mle} and D, as functions of numbers of fragments for different values of p_{gs}. In plot a) interpolated lines are drawn for fragment counts ranging from 15 to 500 in steps of 5 for $p_{gs} = 0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95$, and in plot b) for $p_{gs} = 0.1, 0.5, 0.9$. 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.

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

Fig. 5.3 shows the UPGMA dendrograms for the 5 species, split out for the 3 dissimilarity measures. The dendrograms for $1-D$ and $1-J$ are largely the same. With all 3 dissimilarities the species are separated well. Notice that the $1-D^{mle}$ dissimilarities are closer to 0, as expected. Notice further that the $1-D^{mle}$ dissimilarities are not a simple shift. 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 behaviour 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*.

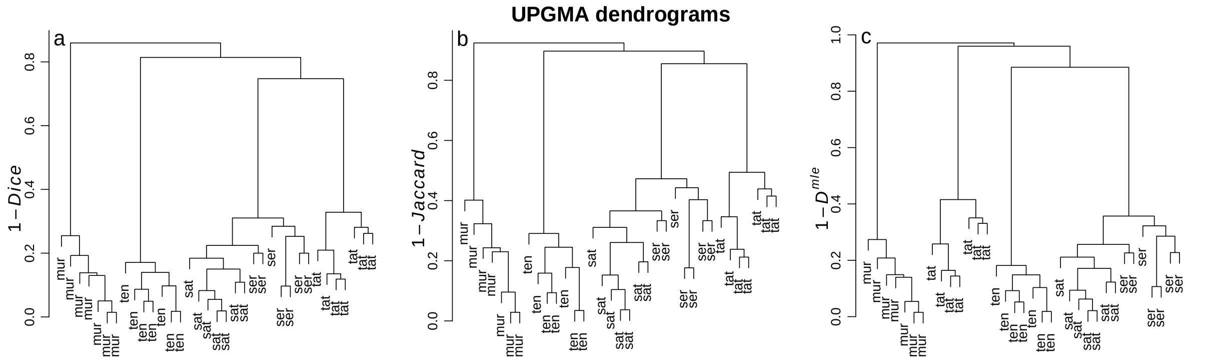


Figure 5.3: UPGMA dendrograms for 3 dissimilarities: a) 1-Dice D, b) 1-Jaccard J , and c) 1-D^{mle} for five species of *Lactuca* and related genera, with 6-9 accessions per species. Labels are: ten = *L. tenerrima*, mur = *M. muralis*, ser = *L. serriola*, sat = *L. sativa*, and tat = *L. tatarica*. For D^{mle} we used *fld* F_S with scoring range 110-501.

5.5 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 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 3 candidate estimators: D^{mle}, D_{L}^{mod}, and $D_{\bar{L}}^{mod}$. Best results were obtained using the maximum likelihood estimator D^{mle}, although differences were small. Second best was, 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} were 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 were 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 $D_{\bar{L}}^{mod}$ and D_{L}^{mod} show more stable coverage properties than those on 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 straightforward for the modified coefficients

$D_{\overline{L}}^{mod}$ and D_{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-monomophic 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 (Eeuwijk & 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 Kraus (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 coancestry, 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, Clarke, and Meudt (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 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.

5.6 Acknowledgements

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.

5.A Appendix Comparison of bootstrap confidence intervals

We compare three types of bootstrap confidence intervals (c.i.):
1. simple percentile c.i.
2. bias-corrected percentile c.i.
3. 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 j-th observation, and $\hat{\Theta}_{.}$ the average of $\hat{\Theta}_{-j}$ ($j = 1, \ldots, N$). In our case the parameter is the fraction of common fragments p_{gs}, estimated by either $D_{\bar{L}}^{mod}$ or D_{L}^{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 $D_{\bar{L}}^{mod}$ 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 $D_{\bar{L}}^{mod}$ 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 $D_{\bar{L}}^{mod}$ 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_L^{mod}, based on a bootstrap resample size of 1000. The results are shown in table 5.5. 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 5.6). 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.

Parameter			Results for $D_{\bar{L}}^{mod}$: 95% bootstrap confidence intervals for p_{gs}					
settings			percentile bootstrap ci		bias corrected bootstrap ci		BC_a ci	
p_{gs}	m_1	m_2	non-cov% (too low, too high)	length(trunc)	non-cov% (too low, too high)	length(trunc)	non-cov% (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
	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
	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

Table 5.5: Comparison of bootstrap confidence intervals for p_{gs} from a simulation study on $D_{\bar{L}}^{mod}$ 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, 1000 bootstrap resamples, $fld\ F_S$ from *A. thaliana* with $N = 450$. 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.

Parameter settings			Results for D_L^{mod}: 95% bootstrap confidence intervals for p_{gs}					
			percentile bootstrap ci		bias corrected bootstrap ci		BC_a ci	
p_{gs}	m_1	m_2	non-cov% (too low, too high)	length(trunc)	non-cov% (too low, too high)	length(trunc)	non-cov% (too low, too high)	length(trunc)
0.0	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
	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
	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

Table 5.6: 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, 1000 bootstrap resamples, $fld\ F_S$ from *A. thaliana* with $N = 450$. 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.

5.B Appendix 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 lane $j = \pi M_j$, 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 lane $j = \sum_{i=1}^{N} y_{ij}$	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 population $j = M_a/M_j$	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
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 (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
$D_{\bar{L}}^{mod}$	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
$J_{\bar{L}}^{mod}$	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}	back-transformed estimator of p_{gs}, using m.l. estimation of $logit(p_{gs})$	stochastic

Chapter 6

Codominant Scoring of AFLP in Association Panels [1]

by Gerrit Gort and Fred A. van Eeuwijk

6.1 Summary

A study on the codominant scoring of AFLP markers in association panels without prior knowledge on genotype probabilities is described. Bands are scored codominantly by fitting normal mixture models to the band intensities, employing the EM-algorithm. We study features that improve the performance of the algorithm, and the unmixing in general, like parameter initialization, restrictions on parameters, data transformation, and outlier removal. Parameter restrictions include equal component variances, equal or nearly equal distances between component means, and mixing probabilities according to Hardy-Weinberg Equilibrium. Histogram visualization of band intensities with superimposed normal densities, and optional classification scores and other grouping information, assists further in the codominant scoring. We find empirical evidence favoring the square root transformation of the band intensity, as was found in segregating populations. Our approach provides posterior genotype probabilities for marker loci. These probabilities can form the basis for association mapping and are more useful than the standard scoring categories A,H,B,C,D. They can also be used to calculate predictors for additive and dominance effects. Diagnostics for data quality of AFLP markers are described: preference for three components mixture model, good separation between component means, and lack of singletons for the component with highest mean. Software has been developed in R, containing the models for normal mixtures with facilitating features, and visualizations. The methods are applied to an association panel in tomato, comprising 1175 polymorphic markers on 94 tomato hybrids, as part of a larger study within the Dutch Center for BioSystem Genomics.

[1] Submitted to *Theoretical and Applied Genetics*

6.2 Introduction

AFLP, or amplified fragment length polymorphism (Vos et al., 1995), is a widely used DNA fingerprinting system. The physical end product of the AFLP procedure is a slab gel, containing bands at different positions within columns of the gel. Instead of gels, capillary systems are nowadays often used. The columns are called lanes, and correspond to the different individual genomes (individuals). The bands visualize amplified DNA fragments of specific lengths, traveling in the lanes by electrophoresis. The position of a band within a lane is mainly determined by the size of the fragment, with shorter fragments traveling further. The pattern of bands within a lane is called a profile. Usually, AFLP bands are scored dominantly, that is, binary, as absent or present. In this way, AFLP bands are dominant markers, which do not distinguish between individuals with one copy of the DNA fragment (heterozygous individuals) and two copies (homozygous individuals). However, the gels or capillary systems allow the intensities of the band to be scored as well. Assuming that the intensity of a band is a measure of the amount of amplified DNA, the band intensity can be exploited to infer the copy number of a DNA fragment. In the case of diploid organisms, an individual with the DNA fragment on two homologous chromosomes (homozygous AA) should have a more intense band than an individual with the DNA fragment on only one of two homologous chromosomes (heterozygous Aa). The heterozygous individual in turn should have a more intense band than an individual, lacking the fragment completely (homozygous absent aa). Therefore, it must be possible to infer the copy number of an AFLP fragment from the band intensity, making the AFLP marker a codominant marker. Scoring the copy number of the AFLP fragment is also named genotype calling.

The idea to codominantly score AFLPs using the band intensities is not new. An early mention can be found in van Eck et al. (1995), and later Piepho and Koch (2000), and, in a reaction, R. C. Jansen et al. (2001) published about the statistical principles of the approach. These authors illustrate their methods by codominantly scoring AFLP markers from segregating F_2 populations, with *a priori* known genotype frequencies 0.25, 0.50, and 0.25 for AA, Aa, and aa, respectively. As Meudt and Clarke (2007) report, codominant AFLP scoring so far is limited to model organisms and commercial crop organisms, for which genetic information already exists for accurate identification of the codominant scores. Vuylsteke (2007) mentions that codominant scoring of AFLP markers has become routine in segregating populations, as in F_2 or backcross populations. Examples of studies of segregating populations, with known segregation ratio for the offspring, are e.g. Castiglioni, Ajmone-Marsan, van Wijk, and Motto (1999), Reamon-Büttner, Schondelmaier, and Jung (1998), and Deniau et al. (2006).

The aim of our study is to extend existing methodology for the codominant scoring of AFLP markers in association panels, without *a priori* knowledge of allele frequencies. The methodology is illustrated using AFLP markers in a collection of 94 tomato hybrids, for which, due to confidentiality reasons, no pedigree information was made available.

An overview of the dataset, and analyses concerning diversity and linkage disequilibrium, containing a concise description of the codominant scoring, can be found

in van Berloo, Zhu, et al. (2008). Commercially available software, like Quantar Pro (Keygene products B.V., 2004), is rather limited in output facilities, as it gives hard classifications only, and does not contain options to back up the codominant scoring in case of an association panel. We therefore developed software ourselves, and used it for the codominant scoring of the AFLP data.
In the present paper we describe

1. the method of codominant scoring of AFLP bands by normal mixture models;
2. some features, that may enhance or stabilize the unmixing of the groups in association panels, where the mixing proportions are unknown in advance;
3. the output from codominant scoring: a) posterior genotype probabilities of the 3 codominant classes, replacing the hard A-B-H-C-D classification which is usually given; b) predictors for additive and dominance effects in QTL analysis calculated from the posterior class probabilities;
4. the dataset, used for illustration of the codominant scoring, consisting of an unstructured association panel of 94 tomato hybrids;
5. the software we developed for the codominant scoring of AFLP profiles in association panels by normal mixture models;
6. an application of the methodology, using the software, to the collection of tomato hybrids.

6.3 Material and Methods

6.3.1 Codominant scoring of AFLP band intensities by normal mixture models

Band intensities
The intensity of an AFLP band, named optical density by Piepho and Koch (2000), is a non-negative number, indicating the darkness of a band on a gray scale. Because band intensities vary from lane to lane (e.g. caused by differences in amount of DNA loaded in a lane), and due to background variation in intensity and image artifacts, the raw band intensities need to be corrected to make bands comparable between lanes. Corrections can be done in different ways. Piepho and Koch (2000) suggest to remove systematic trends discernible from monomorphic bands with the use of a quadratic polynomial regression models and random lane effects, and to check for spatial correlation. In the present study, we use the correction as performed by the proprietary software of Keygene NV. This correction accounts for total lane intensity and intensity of monomorphic bands, and divides intensities per marker by the maximum value, resulting in a range $0 - 1$.

Codominant scoring
The (corrected) band intensity is related to the amount of amplified DNA at the band position. We assume a monotonous relationship: more amplified DNA tends to produce darker bands. This means for diploid organisms, like tomato, that a homozygous individual with two copies of a fragment tends to have a band with higher intensity than a heterozygous individual with a single copy, which in turn has a higher intensity band than an individual lacking the fragment completely. Codominant scoring of a band is the prediction of the copy number of

the fragment (or genotype class AA, Aa, or aa) from the intensity of the band. Codominant scoring is straightforward in the case that the intensities fall into three well-separated groups. But more often, groups overlap, e.g. because the relationship between band intensity and copy number is non-linear, as indicated by Piepho and Koch (2000). The intensity may be upwardly bounded due to saturation, hampering the discrimination between heterozygous and homozygous individuals. Other problems, blurring simple inference on zygosity, are errors in the AFLP procedure itself (like amplification errors in the Polymerase Chain Reaction (PCR), and gel mobility errors), and measurement errors of the band intensities. To take account of these problems, a formal approach using a statistical model is beneficial.

Normal mixture models

Statistically speaking, codominant scoring is a type of cluster analysis with a predefined number of classes (3 in the case of diploid organisms). Although ordinary clustering techniques could be used, the common approach described in the literature is to fit a Gaussian (or normal) mixture model. This is an example of model-based clustering (Fraley & Raftery, 2002), because a proper statistical model is used to describe the data. For an association panel of n individuals, we have per marker n intensities, labeled $y_1 \ldots y_n$. The Gaussian mixture model (G. McLachlan & Peel, 2000) for intensity y_i of variety i is:

$$f(y_i) = \sum_{j=1}^{3} \pi_j f_j(y_i) \tag{6.1}$$

with f_j the density of a normal distribution with mean μ_j and standard deviation σ_j. The mixing probability π_j is the prior probability that a randomly drawn individual belongs to group, or component, j. In the standard situation, we have 3 groups: 1=no copies, 2=one copy, and 3=two copies. We assume for the expected intensities μ_j, that $\mu_1 < \mu_2 < \mu_3$. The posterior probability of cultivar i to belong to group k ($k = 1, 2, 3$) is

$$\tau_{ik} = \frac{\pi_k f_k(y_i)}{\sum_{j=1}^{3} \pi_j f_j(y_i)}, \tag{6.2}$$

which are conditional genotype probabilities given the marker phenotype (intensity). In total 8 unknown parameters are to be estimated: μ_1, μ_2, μ_3, σ_1, σ_2, σ_3, and π_1, π_2 (and $\pi_3 = 1 - \pi_1 - \pi_2$), using maximum likelihood. For segregating populations parameter values may be known. E.g. in case of F_2 populations, the segregation ratio is 1:2:1, hence $\pi_1 = 0.25, \pi_2 = 0.5, \pi_3 = 0.25$. We use the EM-algorithm (Dempster et al., 1977) to get maximum likelihood estimates, treating the situation as an incomplete data problem with missing class memberships, as in R. C. Jansen (1993). In the algorithm, the E-step, in which estimates of the posterior class probabilities $\hat{\tau}_{ik}$ are returned by conditioning on data and parameters, and M-step, returning new parameter estimates $\hat{\mu}_k, \hat{\sigma}_k, \hat{\pi}_k$, alternate until convergence. The M-step consists of separate update steps for π_j, fitting a generalized linear model for multinomial data to the weights $\hat{\tau}_{ik}$, and for μ_j and σ_j, fitting a

linear model allowing for 3 group means (ANOVA model) and weights $\hat{\tau}_{ik}$ to the replicated intensities.
In non-standard situations the number of components g of the normal mixture model may deviate from 3. We refer to item 4a of the next section.

6.3.2 Features for enhanced and stabilized unmixing, data quality and model selection

We study a number of features relevant to the codominant scoring methodology in association panels. Some of them relate to the EM-algorithm, aiming at enhancement or stabilization of the unmixing, others at assessment of the quality of the AFLP marker data for codominant scoring, or model selection.

1. Starting values
   To start up, the EM-algorithm needs either starting values of the parameters (μ_k, σ_k, π_k), followed by an E-step, or starting values of posterior probabilities (τ_{ik}), followed by an M-step. Badly chosen starting values could result in convergence to a local likelihood maximum or non-convergence of the algorithm. We investigate two types of starting values for the EM-algorithm:
   a) guesstimates of the parameters, based on the number of groups (g), and minimum and maximum of the intensities, assuming equidistant $\hat{\mu}_k$, constant $\hat{\sigma}_k = (max - min)/2g$, and constant $\hat{\pi}_k = 1/g$;
   b) cluster based starting values, obtained from a hierarchical cluster analysis (UPGMA), cutting the dendrogram at the desired number of clusters, and calculating means, standard deviations, and relative frequencies within the clusters.
2. Restrictions on parameters
   The modeling principle of parsimony dictates to find models as simple as possible, yet capturing the essence of the data. In our case, putting restrictions on standard deviations, means, and/or prior probabilities may be beneficial.
   a) Standard deviations σ_j
      Models with different standard deviations for the different components tend to produce unstable results, especially if the number of observations in a group is small. Therefore, a model with a single standard deviation, common to all components, is to be preferred. Usually a data transformation is needed to achieve approximate homoscedasticity, see item 3.
   b) Means μ_j
      Assuming a linear relationship between band intensity and copy number, the restriction $\mu_2 - \mu_1 = \mu_3 - \mu_2$, or $\mu_1 - 2\mu_2 + \mu_3 = 0$, may be in place. With this restriction only 2 mean parameters are left. This restriction can be easily built into the mixture model by fitting at the M-step for μ_k not an ANOVA model, but a simple linear regression model with the copy number as regressor. A less stringent restriction, still preventing the means to "go anywhere", penalizes the second order differences between μ's, but needs a smoothing parameter λ to be specified. This leads to penalized weighted least squares at the M-step of the EM-algorithm.
   c) Prior probabilities π_j
      In the codominant scoring of an association panel no knowledge is available

about the prior probabilities π_j. Yet it may be fruitful to restrict the parameters assuming Hardy-Weinberg equilibrium (HWE), as in R. C. Jansen (1994), rendering a single parameter p, representing the allele frequency of the marker in the population. The restrictions on π_j according to HWE are: $\pi_1 = p^2$, $\pi_2 = 2p(1-p)$, $\pi_3 = (1-p)^2$.

3. Allowance for heteroscedasticity
   Band intensities generally show non-constant standard deviation: larger intensities tend to have larger variability. Taking the relationship between variance and mean into consideration, we may arrive at a simpler model with a single dispersion parameter, as described in 2a. This could be done in different ways:
   a) Transformation of band intensity
      R. C. Jansen et al. (2001) mention that band intensities need to be square-root transformed, as this leads to distributions with constant variance. Note however, that this transformation stabilizes the variance only if the variance is proportional to the mean. To allow for other variance-mean relationships, we will study power transformations y^λ, with power λ possibly different from 0.5.
   b) Non-normal mixtures
      Another way to deal with the relationship between variance and mean is to model it directly, allowing a mixture of non-normal distributions. To this end, at the M-step for μ a generalized linear model may be fitted with variance proportional to the mean and log link, using quasi likelihood (McCullagh & Nelder, 1989). We will not pursue this topic further in the results section.
4. Diagnostics for quality of AFLP band intensity data in codominant scoring
   a) Number of groups g
      In case of diploid organisms we assume mixture models with 3 components, allowing for 0, 1, or 2 copies of a DNA fragment. We may, however, face situations with only 2 components, if 0 or 1 copy, 0 or 2 copies, or 1 or 2 copies of a DNA fragment occur in the collection of individuals. Even situations with more than 3 components cannot be ruled out, because collisions may have occurred (Gort et al., 2008). In case of collision two or more different fragments of the same length were amplified for one or more individuals, appearing as single bands. Each fragment may then occur singly (heterozygous) or doubly (homozygous). The band intensity is expected to be highest for the individual with collision. Outliers in band intensity from unknown origin could also cause the number of components to deviate from the expected $g = 3$. The relative goodness of fit of the mixture model with 3 components, compared to models with other numbers of components, will be used as diagnostic for data quality of an AFLP marker for codominant scoring (see also paragraph on Model comparison below).
   b) Separation of groups
      If groups are not well separated, it may be difficult to infer the correct number of groups. Lindsay (1995, pg 18-19) mentions that, for a 2-component normal mixture with means less than two standard deviations apart (corresponding to a unimodal mixture), there is almost no information about the mixing proportion. With a separation of 4 standard deviations or more

the information is almost complete. To check the separation of groups, we propose to calculate for each AFLP marker $sep_1 = (\hat{\mu}_2 - \hat{\mu}_1)/\hat{\sigma}$ and $sep_2 = (\hat{\mu}_3 - \hat{\mu}_2)/\hat{\sigma}$ in the 3-component normal mixture model with constant standard deviation σ. We call the separation "poor" if $sep_1 \leq 2$ or $sep_2 \leq 2$, "moderate" if not "poor", but $2 < sep_1 \leq 4$ or $2 < sep_2 \leq 4$, and "good" if $sep_1 > 4$ and $sep_2 > 4$. The classification of the separation is a second diagnostic for data quality of AFLP markers in codominant scoring.

c) Outliers
For some markers, one or two individuals may have excessively high intensities. Many approaches exist to handle outlying observations in mixture models, like filtering out outliers by addition of a uniformly distributed component to the mixture, or robustification of the procedure using mixtures of t-distributions (see e.g. G. J. McLachlan, Ng, & Bean, 2006). We take a simpler route here, and use two approaches: 1) identify outlying observations by simple visual inspection of the histogram (see item 5), and, if needed, refit the mixture model after removal of these observations; 2) check the number of individuals in the component with highest (and lowest) group mean, according to the classification by the mixture model; if a single observation (singleton) is observed, the band intensity may be outlying. Lack of outliers is a third diagnostic for data quality.

5. Visualization of data and results
As a helpful tool in judging the fit of a mixture model to the data, we use histogram visualization of the band intensities with superimposed density plots, as in R. C. Jansen et al. (2001), and optionally a color-coded hard classification of individuals. Because the mixture model is fitted to corrected intensities (in the range $0-1$, see section 6.3.1), it may be helpful to add as extra information to the histogram the minimum and maximum value of the raw uncorrected intensities (in the range $0- \approx 10^6$), because these reveal relevant information about the gray levels of the bands. Plotting optionally extra grouping information, like tomato type (with levels beef, round, or cherry in the tomato dataset), along the top part of the histogram, may also help the interpretation of the mixture results.

Model comparison
Comparison of nested models is usually done by likelihood ratio tests, but in the case of mixture models theoretical problems of non-identifiability arise. Hypothesis testing in mixture models is a topic of ongoing statistical research (see e.g. Garel, 2007; Chen & Li, 2009). We take special interest in

1. Testing for Hardy-Weinberg Equilibrium
To test the null hypothesis of mixing probabilities according to HWE, we use the likelihood ratio test (LRT), assuming under H_0 a χ^2_1 distribution of the test statistic $LR = 2(LL(FM) - LL(RM))$, with $LL(FM)$ the log-likelihood of the full model with unrestricted π_i, and $LL(RM)$ the log-likelihood of the restricted model with estimated π_i according to HWE. Given the theoretical problems with LRT's in mixture models, we underpin this approach by a small simulation study. We simulate band intensities for 100 individuals, by sampling from a 3-component normal mixture with means $\mu = 0.3, 0.5, 0.7$, a range of

standard deviations $\sigma = 0.025, 0.030, 0.035, 0.040, 0.045, 0.050$, and a range of allele frequencies $p = 0.5, 0.4, 0.3, 0.2, 0.1$. (This set of parameters results in histograms similar to those that occur in the tomato dataset used for illustration, see section 6.3.4.) For the simulation, we first sample the genotype for 100 individuals, using a multinomial distribution with prior probabilities p^2, $2p(1-p)$, and $(1-p)^2$, resulting in counts (k_1, k_2, k_3) representing k_1 homozygous present, k_2 heterozygous, and k_3 homozygous absent genotypes. We do not allow the counts to be zero. Next, we sample k_i intensities from $\mathrm{N}(\mu_i,\sigma)$. From the fitted full and reduced models LR is calculated, and compared to the 95% critical value 3.84 of the χ^2_1 distribution. This procedure is replicated 10.000 times, and type I error rates are calculated.

2. Order selection, i.e. the choice of the number of components of the mixture model. Following Fraley and Raftery (2002), we use the Bayesian Information Criterion $BIC = -2LL + d \times ln(n)$ to compare models with different numbers of groups, where d is the number of parameters, and n is the number of observations. A smaller value of BIC indicates a better fitting model. The "best fitting model" thus corresponds to best fitting according to BIC.

In other cases we compare fits of models by comparing BIC's. If the compared models have equal numbers of parameters, the comparison by BIC is equivalent to the comparison by LL.

6.3.3 Output from codominant scoring

Hard classification versus posterior probabilities

The usual result from the codominant scoring of AFLP markers is a hard classification of markers into categories. The classification can be done in different ways. Piepho and Koch (2000) suggest to take the category with highest posterior probability. The proprietary genotyping software of Keygene NV uses classification rules suggested by R. C. Jansen et al. (2001): genotype i is classified as:

A= homozygous = genotype class AA (= 2 copies), if the posterior probability $\hat{\tau}_{i3} \geq 0.98$;

B= homozygous absent = aa (= 0 copies), if $\hat{\tau}_{i1} \geq 0.98$;

H= heterozygous = Aa (= 1 copy), if $\hat{\tau}_{i2} \geq 0.98$;

C= not homozygous = not AA (= 0 or 1 copy), if none of first three conditions is satisfied, but $\hat{\tau}_{i1} + \hat{\tau}_{i2} \geq 0.98$ for an intensity y_i in the left tail of the normal distribution with mean $\hat{\mu}_2$;

D= not homozygous absent = not aa (= 1 or 2 copies), if none of first three conditions is satisfied, but $\hat{\tau}_{i2} + \hat{\tau}_{i3} \geq 0.98$ for an intensity y_i in the right tail of the normal distribution with mean $\hat{\mu}_2$;

U= missing.

The threshold probability 0.98 is the default value, but other values can be chosen as well. We notice that an extra region of doubt is necessary, because it may happen that genotypes exist, which cannot be classified as A, B, H, C or D. This may occur if the groups are not well separated, so that for some genotypes, $\hat{\tau}_{i1} + \hat{\tau}_{i2} < 0.98$, but also $\hat{\tau}_{i2} + \hat{\tau}_{i3} < 0.98$. The right hand side plot of figure 6.1 shows an example. We call this extra region of doubt Z = unknown, meaning 0, 1, or 2 copies. The left hand side plot shows the classification if probability threshold

0.95 is used. In that case all genotypes can be classified as A, B, H, C, or D. The above mentioned commonly used hard classification has a number of disadvantages. For instance, the classification rule, following from the probability threshold 0.98, is rather arbitrarily chosen. Furthermore, it is not clear how to deal with genotypes, once they are classified into one of the regions of doubt. Therefore, we propose to use instead the set of 3 posterior probabilities $(\hat{\tau}_{i1}, \hat{\tau}_{i2}, \hat{\tau}_{i3})$ as result of the codominant scoring for genotype i. Using this approach, each genotype is allowed to belong to more than one class, with the posterior probabilities indicating the levels of membership to the classes. This type of clustering is called fuzzy clustering, see e.g. Bezdek (1981).

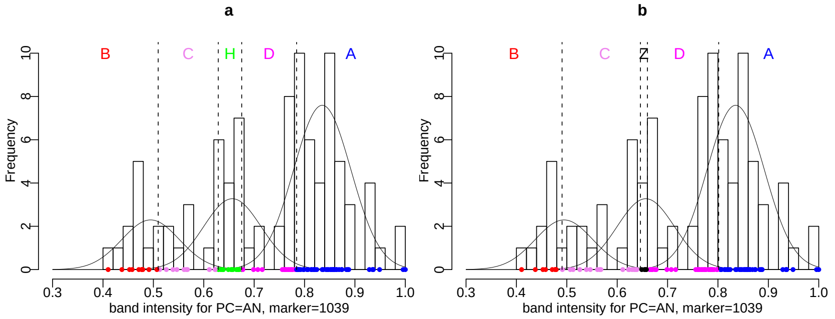


Figure 6.1: Histograms of band intensities of marker 1039 with superimposed normal densities. Subplots a and b show color coded hard classifications based on probability thresholds 0.95, and 0.98 resp. In the last case some observations are classified as unknown (Z).

Predictors for additive and dominance effects
Given the 3 posterior probabilities, it is straightforward to calculate predictors for the additive and dominance effects of the loci. The additive predictor for an individual is defined as $x_a = \hat{\tau}_3 - \hat{\tau}_1$, with values between -1 and 1. The value -1 is obtained for loci which are classified as B (=aa) with probability 1. A locus has additive predictor value 1 if it is classified as A (=AA) with probability one. The dominance predictor x_d depends only on the probability of a heterozygous genotype, and is defined as $x_d = \hat{\tau}_2$, with values between 0 and 1.
The additive and dominance predictors may be used e.g. in QTL-analysis, relating the codominant scores to phenotypic information by mixed models. A paper on genome wide association mapping using these scores is in preparation.

6.3.4 Data: association panel of tomato hybrids

Within the Centre for BioSystems Genomics, a Dutch plant genomic initiative (Berloo, van Heusden, et al., 2008), one project aims at processes and mechanisms affecting fruit quality in tomato. Within this project an association panel, consisting of a diverse set of 94 tomato hybrids, was genotyped using AFLP with gel electrophoresis (Berloo, Zhu, et al., 2008). This set consists of 20 beef, 21 cherry, and 53 round tomato hybrids. The AFLP fingerprinting was performed at Keygene NV using standard in-house developed protocols. Fifty primer combinations were used, labeled A, B, ..., Z,AA,AB,...,AX, based mostly on *Eco*RI / *MSe*I and some *Pst*I / *MSe*I restriction enzyme combinations. The scoring range is approximately 50-550. Typically, between 50 and 100 bands are visible per primer combination per variety, the majority of which is monomorphic. Band intensities of a total of 1175 polymorphic bands were scored by Keygene NV using the proprietary genotyping software. For 378 bands the map position is available from an integrated proprietary linkage map. We study both raw uncorrected intensities, with values in the range $0- \approx 10^6$, and corrected intensities with values in the range $0-1$. We refer to the dataset of band intensities of 1175 AFLP markers on 94 tomato hybrids as the "tomato data".

6.3.5 Studying the scoring features in the complete tomato dataset

We study how the features mentioned in section 6.3.2 help in the codominant scoring of all 1175 AFLP markers in the tomato data, focusing on the following topics.

1. Starting values of parameters. We study the performance of the two types of parameter initialization for the EM-algorithm. For each marker, mixture models with 2, 3, 4 and 5 components are fitted, once using guesstimates and once using cluster based starting values. We tabulate how often each type of starting values performs best (highest LL).
2. Power transformation of the band intensity. We try to find empirical evidence favoring the square root transformation, as suggested by R. C. Jansen et al. (2001), in two ways:
   a) Comparing the fits of homoscedastic and heteroscedastic 3-component mixture models for power transformations in the range 0.25-1.0 with BIC. Per transformation we count how often the homoscedastic model (with $d = 6$ parameters) is preferred over the heteroscedastic model (with $d = 8$). If the estimated standard deviation $\hat{\sigma}$ in a mixture component is smaller than 0.01, or if a component contains a single observation, we fix $\hat{\sigma}$ at 0.01. The power transformation, giving most often variance stabilization, is called best with respect to variance.
   b) Comparing the fits of mixture models with 2, 3, 4, and 5 components for power transformations in the range 0.25-1.0, using BIC. Per power transformation and marker, the best fitting model is selected. The transformation, selecting most often the preferred 3-component mixture model, is called best with respect to order selection.
3. Diagnostics for data quality of the 1175 AFLP markers:
   a) number of components: compare g-components homoscedastic mixture models (with $g = 2, 3, 4, 5$ components, and $d = 4, 6, 8, 10$ parameters, resp.) by BIC;
   b) separation: count how often separation is poor, moderate or good in the best-fitting g-components model;
   c) outliers: count how often singletons exist in the first or last component in

the best-fitting g-components model.

4. Hardy-Weinberg Equilibrium. We test the null hypothesis of mixing probabilities according to HWE for a subset of markers, using the LRT described in section 6.3.2. We use a selection of 300 mapped markers, following the paper by Berloo, Zhu, et al. (2008). Out of the 797 unmapped markers, we select 349 with best fitting 3-components mixture model.

6.4 Results

6.4.1 Software

We developed software routines in R (Ihaka & Gentleman, 1996) for the codominant scoring of AFLP band intensities in an association panel, using the EM-algorithm. We built features into the software, as described in section 6.3, allowing for different starting values of parameters, transformation of the response, restriction on parameters, different numbers of components, and for the types of output as described earlier. For a more detailed description of the software we refer to appendix 6.A. All plots and mixture model output in this paper are results from applications of the R routines.

6.4.2 Examples

Examples with well fitting mixture models
In figure 6.2 we show some examples of codominantly scored AFLP markers with well fitting 3-component homoscedastic normal mixture models. The corrected band intensities are square-root transformed, unless mentioned otherwise. In subplots a) and b) no variety is classified into a region of doubt. In subplots c) and d) a few hybrids are classified as “D”. We added the boundaries of the classes into the plot, and minimum and maximum value of the raw band intensities. The variety in plot c) classified as “D” has posterior probabilities $(\hat{\tau}_{i1}, \hat{\tau}_{i2}, \hat{\tau}_{i3}) = (0, 0.050, 0.950)$.

Examples of features helping unmixing
Figure 6.3 illustrates problems encountered in the codominant scoring of AFLP band intensities of the tomato dataset, that can be handled with the features described in section 6.3.2. The subplots are labeled accordingly.

1. Starting values. Subplots 1a) and 1b) show an example where cluster initialization of the parameters in the EM-algorithm results in a better solution ($LL = 120.1$) than initialization by guesstimates ($LL = 109.1$).
2. Restrictions on parameters.
   a) Standard deviation σ_j. In subplots 2a1) and 2a2) an example of the differences in fit between models with free and equal standard deviations is given. The rather outlying observation is accommodated in subplot 2a1) by allowing for a mixture component with a very large standard deviation. Although the model with free σ_j (with $d = 8$ parameters versus $d = 6$ for the homoscedastic model) has a substantially higher LL (76.6 vs 70.5), resulting in a smaller BIC (-116.9 vs -113.7), visual inspection shows that

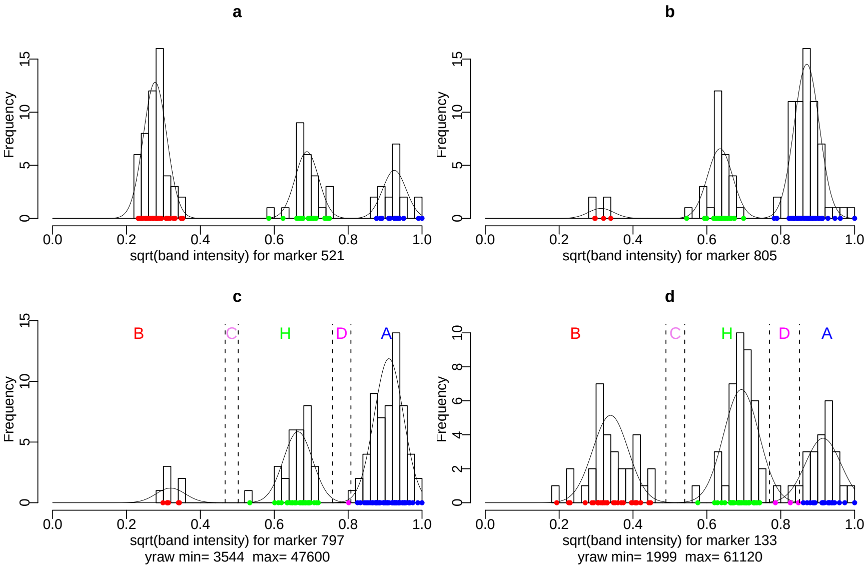


Figure 6.2: Four examples of AFLP markers from the tomato data with histograms of band intensities, and well fitting normal mixture densities.

the restricted model has a more reasonable fit.

b) Means μ_j. For the marker in subplots 2b1) and 2b2) the equidistance restriction on μ_j results in a better solution ($LL = 31.2$) than the model with free μ_j's ($LL = 21.9$). This is an example of a pathological situation, because the EM-algorithm converges to an inferior solution for the full model (free μ_j's) compared to the restricted (equidistant) model, whereas by definition the larger model must fit better.

c) Prior probabilities π_j. In subplots 2c1) and 2c2) an example is shown, where the model with restricted π_j according to HWE ($\pi_1 = p^2$, $\pi_2 = 2p(1-p)$, $\pi_3 = (1-p)^2$) results in a higher LL (46.8), than the model with free π_j ($LL = 46.0$). Again, the reason must be convergence of the EM-algorithm to an inferior solution for the model with free π_j, in this case by allowing a separate component with small mixing probability for the two hybrids with very low band intensity.

3. Transformation of band intensity. Subplots 3a1) to 3a4) show the interplay between data transformation and restriction on σ_j. In 3a1) and 3a2) mixture models are fitted for untransformed band intensities. The heteroscedasticity has to be taken care of by allowing for different σ_j's. In 3a3) and 3a4) the same AFLP marker is studied, but now the band intensities are square root transformed. For the square root transformed intensities, the simpler model with equal σ_j's is reasonable.

4. Diagnostics for quality of AFLP band intensity data.

   a) Number of groups. Subplots 4a1) and 4a2) show an example with a better fitting 4-component mixture, compared to 3 components, according to BIC.

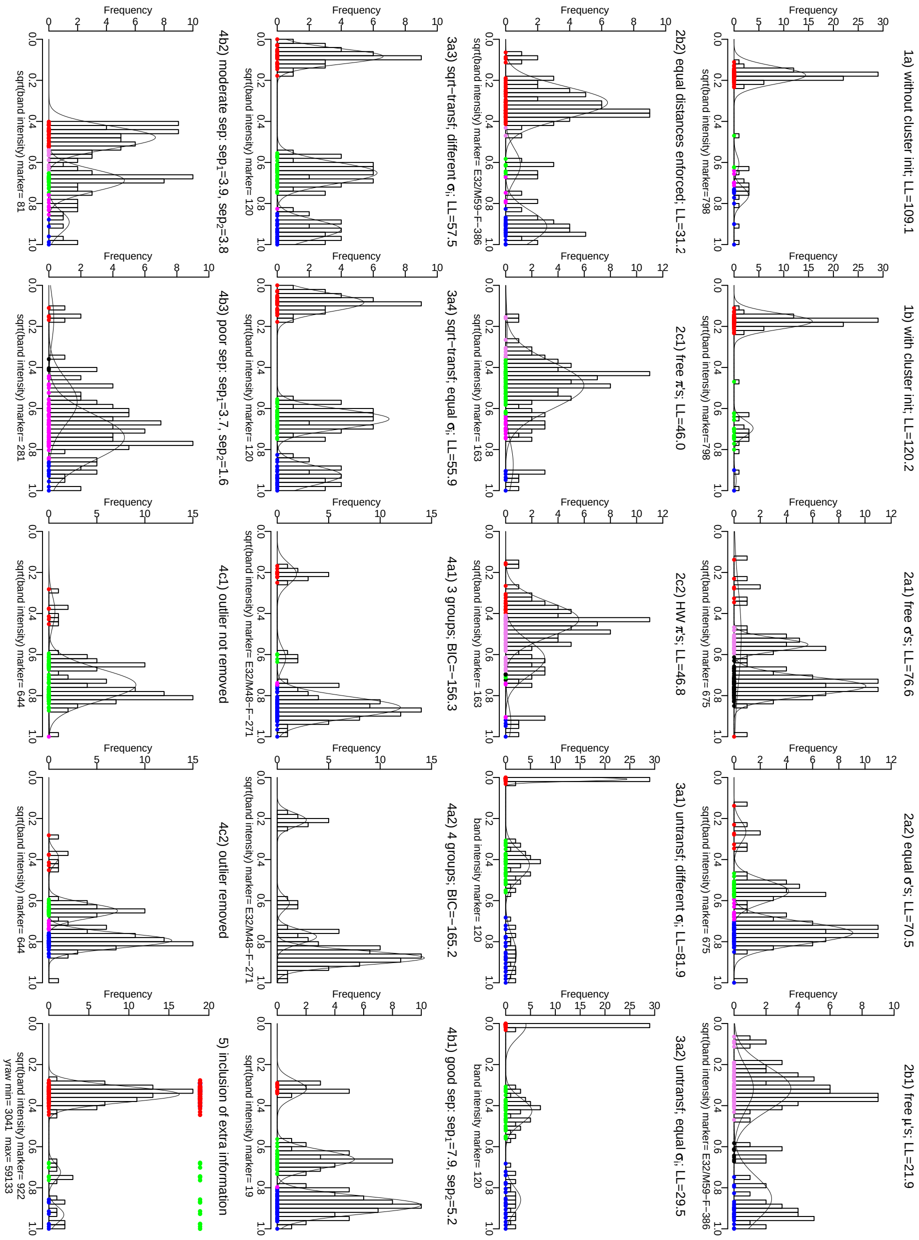


Figure 6.3: Examples of features helping unmixing of marker intensities for the tomato data. Subplots 1a-b deal with starting values of parameters; 2a1-a2 restriction on σ: hetero- vs homoscedasticity; 2b1-b2 restriction on μ: equidistant component means; 2c1-c2 HWE restriction on π; 3a1-a4 transformation of band intensity; 4a1-a2 number of components of mixture model; 4b1-b3 separation of group means; 4c1-c2 outliers; 5 extra information in plot.

b) Separation. Three examples of markers with good, moderate, and poor separation are shown in subplots 4b1), 4b2), and 4b3). In all three cases the separation between the Aa and AA is worse than between aa and Aa.

c) Outliers. Subplots 4c1) and 4c2) show the effect of removal of an outlier. A separate component of the mixture is devoted to the outlier, if included. Without the outlier the mixing probabilities are nicely according to HWE.

5. Data visualization. In subplot 5) we include extra information: minimum and maximum of the raw intensities, and values of an extra grouping variable, in this case type of tomato, shown as colored dots along the top of the graph. The AFLP marker indicates population substructure, because it is related to tomato type: all genotypes with high intensities are cherry tomatoes (shown as green colored dots).

6.4.3 Results for the complete tomato dataset

Parameter initialization

Table 6.1 shows the comparisons of the two types of parameter initialization of the EM-algorithm (by guesstimates and hierarchical clustering) for 2-, 3-, 4-, and 5-component homoscedastic mixture models for all 1175 markers. We find that parameter initialization becomes more critical for more complex models. In case of mixture models with 2 groups, initialization by guesstimates and by hierarchical clustering results in identical parameter estimates (with maximized log-likelihood differing less than 10^{-6}) for 95% of the markers. For models with 3, 4 and 5 groups this percentage is 74%, 55%, and 34% respectively. For models with more than 2 groups the cluster initialization outperforms the guesstimates. We conclude that cluster initialization is a better procedure for supplying starting values for parameters. To avoid being trapped in a local maximum, however, we advise to try other starting values as well, using e.g. the described guesstimates. In the following analyses we fit models using both types of parameter initialization, and choose the results corresponding to the model with highest LL.

	number of groups			
	2	3	4	5
no difference	1118	870	651	405
guesstimate best	30	73	92	142
cluster best	27	232	432	628
total	1175	1175	1175	1175

Table 6.1: Comparison of parameter initialization by log-likelihood of fitted models: guesstimates versus hierarchical clustering

Transformation of band intensity

Table 6.2 shows the comparison of homoscedastic and heteroscedastic 3-component mixture models by BIC for a range of power transformations. Between 3 and 15 markers, depending upon the transformation used, are discarded, because the LL of the heteroscedastic model is erroneously lower than that of the (smaller) homoscedastic model, due to convergence to local minima. Among the different power transformations, the square root transformation gives most often (63%) variance stabilization.

Power transformation									
0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	1.0
57%	59%	61%	63%	58%	49%	45%	40%	32%	27%

Table 6.2: Comparison of homoscedastic and heteroscedastic 3-component mixture models by *BIC* for a range of power transformations of band intensities. Shown are percentages of markers with the homoscedastic model selected as best.

Table 6.3 shows the results of the comparisons of 2-, 3-, 4-, and 5-component homoscedastic mixture models for a range of power transformations. We find some very distinctive patterns. If the square root transformation is used, the 3-component model is selected most frequently (for 561 markers). Transformation by power 0.6 shows almost similar results. With powers larger than 0.5, models with more groups tend to be favored, probably because large observations tend to become more outlying, which are accommodated by more components. Using a transformation with a power smaller than 0.5, both models with 2, and with 4 or 5 groups tend to be selected more often. We conclude from tables 6.2 and 6.3 that the square root transformation is best, both for variance stabilization and for order selection.

	Power transformation									
g	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	1.0
2	202	197	172	158	147	132	127	122	117	107
3	458	472	505	561	557	517	476	434	357	315
4	334	334	348	332	310	295	308	313	301	261
5	181	172	150	124	161	231	264	306	400	492
total	1175	1175	1175	1175	1175	1175	1175	1175	1175	1175

Table 6.3: Model selection of g-component mixtures models by *BIC* for a range of power transformations. For each power transformation, the numbers of markers out of 1175 are shown with a g-components normal mixture model ($g = 2, 3, 4, 5$ selected as best).

Diagnostics of data quality

Table 6.4 shows results for the diagnostics of data quality. In the comparison of normal mixture models with 2, 3, 4 and 5 components by *BIC*, we find that the desired model with 3 components fits best for 561 markers ($\approx$ 50%). For 158 markers a model with 2 components fits best. Models with more than 3 components are chosen for 456 markers.

Results on the separation of group means in the best-fitting g-components model are shown in the middle part of table 6.4. Notice that the majority of the markers

(69%) have well separated group means, 31% is moderately separated, and only 1 marker is poorly separated. The percentages well separated markers monotonically decrease with the order g of the model: 89%, 80%, 53%, and 34%, resp. We conclude that the separation of group means shows a relationship with the choice of best fitting model.

The bottom part of table 6.4 shows counts of markers with singletons in the last and first component of the best fitting g-component mixture model ($g = 2, 3, 4, 5$). We find that 62 (5%) of the markers have a first component with a singleton. This percentage is not heavily dependent on which model fits best. However, the counts of markers with a singleton in the last component are much higher, and now we do see a clear relationship with best fitting model: for markers with a best fitting 3-component model, only 42 (7.5%) have a singleton in the last component, whereas markers with best fitting 2-, 4-, and 5-component mixture models have singletons in 25%, 26%, and 36% of the cases.

The problem with outlying observations is that they may be, but not necessarily are, erroneous: a component with a singleton may represent a true genotypic situation. If we assume that rare genotypes AA and aa occur approximately equally often across all markers, and that most singletons in the first component represent true aa genotypes, we conclude that if markers with best fitting 3-component mixture model have singletons in the last component, most of these represent true AA genotypes. The much higher percentages of singletons in the last component found for markers with 2-, 4- or 5-component models, suggest that the intensity is erroneous outlying (whatever the reason may be), and need further examination.

	number of components				
	2	3	4	5	total
selected as best	158	561	332	124	1175
poor separation	1	0	0	0	1
moderate separation	17	113	157	82	369
good separation	140	448	175	42	805
singleton in first component	7	24	19	12	62
singleton in last component	39	42	85	44	210

Table 6.4: Diagnostics for data quality: counts of markers with best fitting mixture models with 2, 3, 4, or 5 components using BIC, counts of markers with poor, moderate, or good separation of group means, split with respect to model choice according to BIC, and counts of markers with singletons in the first or last component of the best fitting mixture model.

Testing for mixing probabilities according to Hardy-Weinberg Equilibrium

Table 6.5 shows the results of the simulation study to underpin the LRT for HWE, as described in section 6.3.2. We note that for allele frequencies $p = 0.3, 0.4, 0.5$ the type I error rates are close to the nominal value 0.05. For smaller values of p the LRT is slightly conservative, rejecting the null hypothesis not often enough (with error rates between 0.034 and 0.045). We suspect that the reason is data sparseness: if p is small, $\pi_1 = p^2$ is close to zero, rendering frequently mixtures with only 1 or 2 observations for the first component. We conclude that the LRT

is justified to test for mixing probabilities according to HWE.

	Allele frequency p				
σ	0.5	0.4	0.3	0.2	0.1
0.025	0.052	0.052	0.055	0.045	0.034
0.030	0.054	0.048	0.054	0.043	0.035
0.035	0.052	0.050	0.052	0.043	0.036
0.040	0.053	0.051	0.047	0.039	0.040
0.045	0.053	0.053	0.049	0.038	0.041
0.050	0.052	0.051	0.049	0.038	0.044

Table 6.5: Type I error rate of the likelihood ratio test for the null hypothesis of mixing probabilities according to HWE ($\alpha = 0.05$) for simulated intensities of 100 genotypes using a 3-components normal mixture model with means 0.3, 0.5, 0.7, using 10.000 replicates.

Figure 6.4 shows an example of a marker with mixing probabilities according to HWE. First a mixture model with unrestricted π_j is fitted, shown in subplot 4a), with $LL = 94.2$. Second, a mixture model with π_j according to HWE is fitted, shown in 4b), with $LL = 93.8$ and estimated allele frequency $\hat{p} = 0.78$. The hypothesis test of π_j according to HWE uses the test statistic $LR = 2 \times (94.2 - 93.8) = 0.8$, and has P-value $P(\chi^2_1 \geq 0.8) = 0.37$. Hence, the null hypothesis of HWE is not rejected.

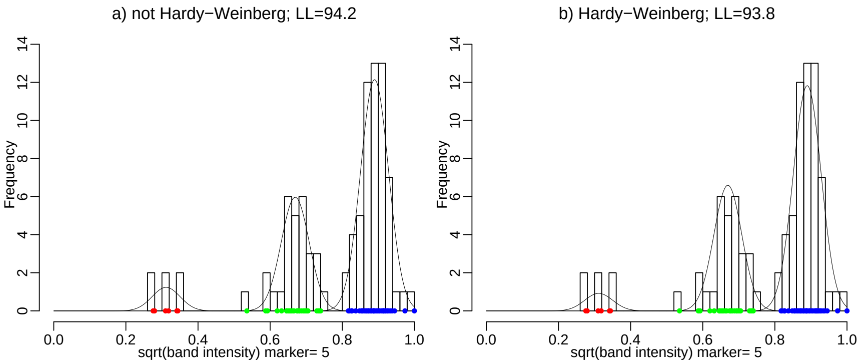


Figure 6.4: Histogram and fitted normal mixtures with unrestricted π_j (subplot a) and restricted π_j according to HWE (b).

The results for all selected markers are shown in table 6.6 (cf table 2 in Berloo, Zhu, et al. (2008)). If the LRT gives a P-value > 0.05, the null hypothesis of HWE for the marker is not rejected, and we accept the mixture model with mixing probabilities according to HWE. We find large differences in percentages of markers in HWE over the chromosomes, with low percentages on chromosomes 4, 5, and 8, to (almost) 100% on chromosome 3 and 9. In the selection of unmapped markers 53% does not show evidence against HWE.

chromosome	1	2	3	4	5	6	7	8	9	10	11	12	unmapped
nr markers	14	5	3	34	28	44	6	7	120	6	19	14	349
nr in HWE	4	3	3	2	5	42	4	1	114	4	10	11	184

Table 6.6: Total numbers of markers and numbers of markers with mixing probabilities according to HWE for a selection of mapped markers on the 12 chromosomes, and of unmapped markers.

6.5 Conclusions and discussion

In this paper we describe a method for the codominant scoring of AFLP markers in association panels without prior knowledge of genotype probabilities. AFLP bands are scored codominantly by fitting normal mixture models to the band intensities per marker, using the EM-algorithm. The EM-algorithm is used for maximum likelihood estimation of normal mixture parameters. It is known for its slow convergence rate, but proved fast enough for the size of the example dataset we analyze here. We study a number of features that facilitate the codominant scoring of AFLP bands, like different parameter initializations for the normal mixture fitting, restrictions on parameters (equal standard deviations, equal or nearly equal distances between component means, mixing probabilities according to HWE), easy data transformation, and outlier removal. Histogram visualization with superimposed normal densities, and optional classification scores and other grouping information assists further in the codominant scoring of the bands. The methods for codominant scoring with facilitating features are implemented in a program in R, that is available from the authors.

Traditionally, the output from codominant scoring based on mixture models is the "hard" classification of genotypes into categories "A","H","B", augmented with regions of doubt "C" (="not A") and "D" (="not B"), for which an extra region of doubt "Z" (="B or H or A") is needed for completeness. It remains unclear how cultivars classified into regions of doubt should be dealt with in further analysis. We therefore propose to replace the hard classification by a fuzzy classification: use the posterior probabilities of individuals to belong to each of the three genotype classes AA, Aa, or aa. The posterior probabilities are direct results of the fitted mixture model without the intervening threshold needed for a hard classification. Given the posterior genotype probabilities, predictors of additive or dominance effects are easy to calculate, and can be used e.g. in association studies.

The EM-algorithm for fitting normal mixture models needs starting values of the parameters. We have studied two types of starting values, and find that cluster based starting values outperform (what we call) guesstimates of the starting values, especially for more complex models. We recommend to fit models twice using both methods for starting values, and choose the fitted model with highest LL.

We find empirical evidence favoring the square root transformation to arrive at homoscedastic normal mixture models.

We have studied criteria for data quality of AFLP markers with respect to codominant scoring, focusing on optimal number of components of the mixture model, separation of components, and occurrence of outliers. In our example dataset (an association panel of tomato), the desired normal mixture model with 3 components, valid for diploid organisms, is selected by BIC for about half of the 1175 polymorphic bands (if choosing from models with 2, 3, 4, or 5 components). A model with more than 3 components is selected for about 38% of the markers. Models with more than 3 components make no sense for diploid organisms, if the components of the mixture model correspond to copy numbers of a unique DNA fragment for the different genotypes. However, if an AFLP band would consist of two different DNA fragments of equal length, which we call collision (see Gort et al., 2006, 2008), a 4 or 5-components model cannot be ruled out. A model with 2 components, which could have a biologically sound interpretation, is selected by BIC for only 13% of the markers.

In total 69% of the markers with best-fitting g-components models have well separated components. This percentage declines with g. Models with good separation are to be preferred, because they will lead to crisp classifications: posterior probabilities close to 0 or 1.

Markers with best fitting 2-, 4-, or 5-components models have in 25-35% of the cases a single observation assigned to the component with highest mean, whereas for markers with best fitting 3-components model this is only 7%. For the component with lowest mean we find 5-10% singletons in all cases. From this, we cautiously conclude that markers, with 2-, 4- or 5-component mixture models selected as best, contain more often an erroneous outlying observation than markers with 3-components models selected best.

From the above we can distill a recipee for the automatic selection of AFLP markers, which can be reliably and consistently scored: select markers with best fitting 3-components mixture model according to BIC, good separation of components, lack of singletons, robustness against parameter initialization, and robustness against slight data transformation.

The LRT to test for mixing probabilities according to HWE appears to be reasonable, as we find from a simulation study. In the example association panel, large differences in percentages of markers in HWE are found between the chromosomes, with percentages ranging from 5% (chromosome 4) to 95-100% (chromosomes 3, 6, and 9). These differences may be caused by recent breeding efforts in tomato focusing on chromosome 4 (Berloo, Zhu, et al., 2008).

For completeness, we note that AFLP markers can be codominant in another sense. If two AFLP fragments differ in size by a few basepairs, e.g. by an indel, but are identical in other respects, and originate from the same locus, they can be used as codominant markers. Such bands or fragments are called allelic markers. Special algorithms and software can find such markers, and score them codominantly (Meudt & Clarke, 2007). An example of a study of this type of codominance is Wong et al. (2007).

Zhanjiang (2007) urges caution in the use of codominant scoring because of the nonlinear nature of the Polymerase Chain Reaction, which is at the basis of the

AFLP procedure, and even discourages the use in case of samples from random mating populations. We have demonstrated, though, in this study of an unstructured association panel of hybrids, that large numbers of AFLP markers can be scored codominantly in a satisfactory way. The main advantage of codominantly scoring AFLPs is obviously being able to distinguish heterozygous from homozygous individuals. Even if some uncertainty about the true genotypic class of a cultivar remains, and some AFLP bands are lost due to low data quality, this advantage makes the codominant scoring of AFLPs in association panels worthwhile.

6.6 Acknowledgements

We thank Ralph van Berloo for a description of the codominant scoring done by Keygene NV. We further acknowledge the contributions of the Center for BioSystems Genomics (project CBSG 2012: BB9-12), and the Generation Challenge Program (project GCP-G4007.09).

6.A Appendix Software description

We wrote software routines for the codominant scoring of AFLP profiles in R (Ihaka & Gentleman, 1996), which are available from the authors. In the software we fit and visualize mixture models, using the EM-algorithm. The main routine takes, besides the normalized intensities and optionally the raw intensities, a number of arguments to allow for the different features described earlier. The arguments are concisely described below.

argument	default	description
`ng`	`=3`	number of groups
`modeltype`	`=2`	`1`= free π, free σ `2`= free π, constant σ `3`= fixed π, free σ `4`= fixed π, constant σ `5`= Hardy Weinberg, free σ `6`= Hardy Weinberg, constant σ
`clust`	`=TRUE`	is clustering initialization of parameters used?
`Pois`	`=FALSE`	is quasi-Poisson regression used to fit models?
`p`	`=1/ng`	starting values and/or fixed values of prior probabilities π_i
`equaldist`	`=FALSE`	are means restricted to be equidistant?
`lambda`	`=0`	value of the smoothing parameter in case of restriction on means
`boxcox`	`=0.5`	transformation of intensities, default is square root
`rm.max`	`=0`	the number of outlying observations that should be removed before unmixing
`pthresh`	`=0.98`	threshold of τ for regions of doubt
`plothist`	`=TRUE`	should a histogram be plotted?

`xlim`	`=c(0,1)`	range of values for x-axis of histogram
`plotscores`	`=TRUE`	should class scores be plotted?
`plotbound`	`=TRUE`	should class boundaries be plotted?
`freq`	`=TRUE`	histogram shows frequencies or densities?
`nbreaks`	`=NULL`	number of classes for histogram
`maintitle`	`=NULL`	the title of the histogram
`showminmax`	`=TRUE`	print minimum and maximum of raw intensities as subtitle
`xlabel`	`=NULL`	extra label at the x-axis
`extrainfo`	`=NULL`	color coded extra grouping information plotted along the top of the plot

The definition of the R function `CodomAFLP` with all arguments follows here:

```
CodomAFLP <- function(y, yraw=NULL, ng=3, modeltype=2, clus=TRUE, Pois=FALSE,
   p=rep(1/ng,ng), equaldist=FALSE, lambda=0, boxcox=0.5, rm.max=0,
   pthresh=0.98, plothist=TRUE, xlim=c(0,1), plotscores=TRUE, plotbound=FALSE,
   freq=TRUE, nbreaks=40, maintitle=NULL, showminmax=FALSE, xlabel=NULL,
   extrainfo=NULL)
```

Routine `CodomAFLP` returns the estimated means, standard deviations, prior probabilities, and posterior probabilities. For mixtures of 2 or 3 groups also the hard classifications are given. In case of Gaussian mixtures the log likelihood is returned as well. Based on the data and the model fit, a histogram visualization with fitted densities can be produced. Optionally, the observations can be plotted on the x-axis using a color coding corresponding to the hard classification. We use the following color codes: red=B, green=H, blue=B, violet=C, magenta=D, black=Z.

Chapter 7

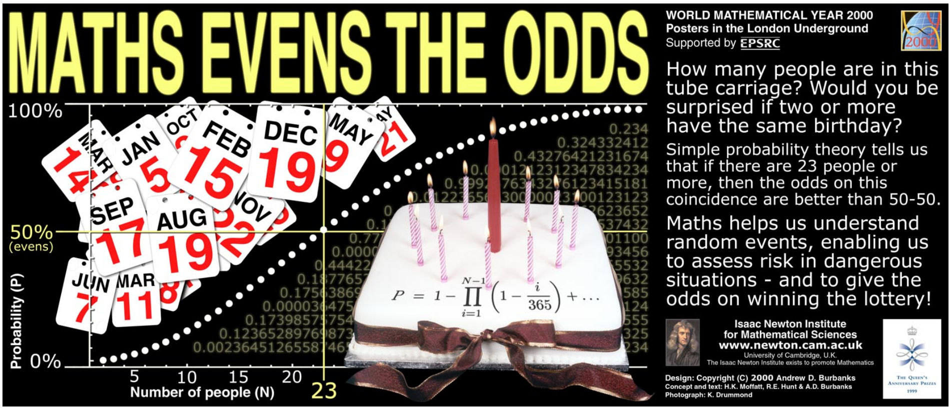
General Discussion

7.1 Introduction

In this thesis we have described some studies on statistical properties of AFLP. These studies were born from a practical question: if similarity between individuals is calculated from AFLP profiles, which values would indicate true phylogenetic relationship? We addressed this question in chapter 2, and proposed a Monte Carlo approach to simulate the distribution of similarity coefficients for unrelated individuals. From this distribution critical values for testing the unrelatedness of individuals could be determined. We also suggested weighted similarity coefficients. After gaining more insight into collision from a probabilistic point of view as described in chapters 3 and 4, a better answer was given in chapter 5. In this chapter we defined modifications of similarity coefficients, that automatically correct for homoplasy and collision. In chapter 6 we studied another aspect of AFLP that relates to collisions: codominant scoring. In that chapter we already touched upon this relationship, but in the present discussion chapter we will take it a bit further in section 7.5.

We find some of our results surprising, in the same way as the birthday problem is surprising: the size of the problem is larger than one would believe at first thought. In a relative small group, the probability that two or more people share a birthday is higher than people tend to believe. Likewise, in an AFLP profile with relatively few fragments, more fragments have equal length, and hence cluster together within a single band, than one may think. Hence the title of the thesis: on some surprising statistical properties of AFLP. The surprising aspect is clear from the poster (see figure 7.1) that could be found in the London underground in the year 2000, being one of a series of 12 monthly mathematical posters, which were meant to raise awareness of the importance of mathematics among the broad public. A second poster about the Human Genome project, also related to our work when we considered the complete Arabidopsis genome in chapter 2, is shown in figure 7.2. The posters were one of the expressions of a campaign sponsored by the Isaac Newton Institute in Cambridge, following Unesco's decision to label the year 2000 as World Mathematical Year. The posters showed the public that mathematics has an enormous range of important practical applications. As Keith Moffatt, the then president of the Isaac Newton Institute, stated: “We are trying to bring it to life, to show people it is not just this awful business of numbers and sets and diagrams.” (The Guardian, Dec 30th, 1999).

In this discussion chapter we look back at our findings, and relate them to the findings of others that we compiled in chapter 1. We will make some critical remarks with respect to our own work. Next, we sketch some potential future work. Finally, we wrap things up, and reach our final verdict.



WORLD MATHEMATICAL YEAR 2000
Posters in the London Underground
Supported by EPSRC

How many people are in this tube carriage? Would you be surprised if two or more have the same birthday?

Simple probability theory tells us that if there are 23 people or more, then the odds on this coincidence are better than 50-50.

Maths helps us understand random events, enabling us to assess risk in dangerous situations - and to give the odds on winning the lottery!

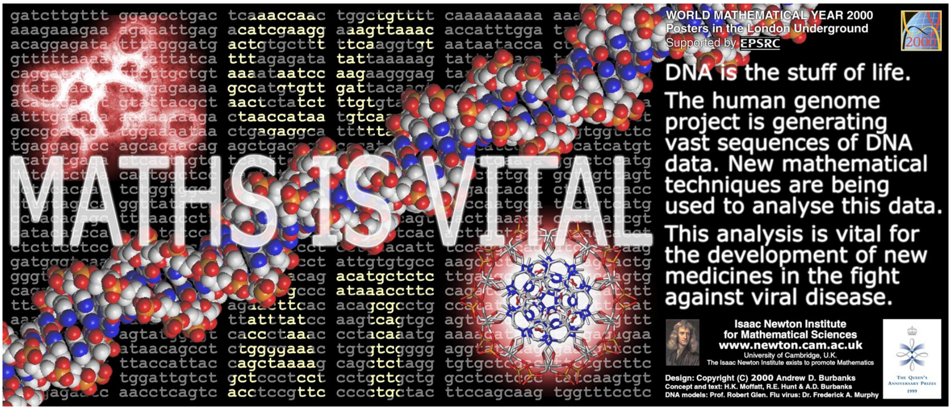
Isaac Newton Institute
for Mathematical Sciences
www.newton.cam.ac.uk
University of Cambridge, U.K.
The Isaac Newton Institute exists to promote Mathematics

Design: Copyright (C) 2000 Andrew D. Burbanks
Concept and text: H.K. Moffatt, R.E. Hunt & A.D. Burbanks
Photograph: K. Drummond

Figure 7.1: Poster on birthday problem in the London underground (source: http://www.newton.ac.uk/wmy2kposters, ©Isaac Newton Institute for Mathematical Sciences, Cambridge, UK).

7.2 Theory and empiricism

We started this thesis in section 1.4 with a compilation of papers that study homoplasy and collision. These were all empirical studies, focussing more on cases than on methods. In contrast, our work emphasizes methods, with theory and statistics as driving forces. What have we learned? How do our results compare to the findings of others? In this section we make a comparison between the results reported by the mentioned authors and the results we would get given their data. The problem with this approach is, that often some necessary information is missing in the papers. For example, to estimate the number of collisions within a lane, we need the total number of bands within a lane or the lengths of all bands within a lane, and the fragment length distribution or at least the GC content of the genome involved. In some cases we impute the missing information, but this may make a fair comparison difficult. In other cases we don’t even try to fill the gaps, but just comment on the findings by the authors in the light of our results.



WORLD MATHEMATICAL YEAR 2000
Posters in the London Underground
Supported by EPSRC

DNA is the stuff of life.

The human genome project is generating vast sequences of DNA data. New mathematical techniques are being used to analyse this data.

This analysis is vital for the development of new medicines in the fight against viral disease.

Isaac Newton Institute
for Mathematical Sciences
www.newton.cam.ac.uk
University of Cambridge, U.K.
The Isaac Newton Institute exists to promote Mathematics

Design: Copyright (C) 2000 Andrew D. Burbanks
Concept and text: H.K. Moffatt, R.E. Hunt & A.D. Burbanks
DNA models: Prof. Robert Glen. Flu virus: Dr. Frederick A. Murphy

The Queen's Anniversary Prizes 1999

Figure 7.2: Poster on the human genome project in the London underground (source: http://www.newton.ac.uk/wmy2kposters, ©Isaac Newton Institute for Mathematical Sciences, Cambridge, UK).

1. Monte Carlo and in-silico AFLP studies
   a) Vekemans et al. (2002) report for *P. lunatus* 83 collisions in 250 fragments (33.2%) and ≈ 34.7 collisions in 150 fragments (23.1%), and for *L. perenne* 60 collisions in 220 fragments (27.3%) and ≈ 9.8 collisions in 80 fragments (12.3%). They use fld's based on GC content 0.45% for *P. lunatus* and 0.50% for *L. perenne*, and a scoring range of $75-450$. Using methods described in sections 3.4.1 and 3.5.1, we get the following results: for *P. lunatus* 82.9 ± 5.6 collisions in case of 250 fragments, and 33.7 ± 4.3 collisions in case of 150 fragments (cf. 83 and ≈ 34.7 mentioned above); for *L. perenne* 60.1 ± 5.1 collisions for 220 fragments, and 9.2 ± 2.6 collisions in case of 80 fragments (cf. 60 and ≈ 9.8 mentioned above). These results are strikingly similar. This should not surprise us too much, though, because the authors sample AFLP data using the fld, according to Innan et al. (1999), which we can use in our calculations as well. Furthermore, they report averages over many simulated profiles, resulting in expected numbers of collisions, as we do.
   b) Althoff et al. (2007) uses in-silico AFLP on sequenced genomes from eight organisms, from bacteria to humans, representing a range of genome sizes. The authors report total numbers of bands per profile, and numbers of bands containing collisions, split into collisions within and among chromosomes. Unfortunately, this information is not enough to derive the total number of bands with collisions, because a band containing more than one collision could have collisions of fragments within the same chromosome, and with other fragments on another chromosome. This band would be counted twice, once within chromosomes and once among chromosomes. Therefore, the total number of bands with collisions is *at most* the number reported. On the other hand, no information is given about the number of collisions per band, and, especially in cases of large band counts, double, and higher order collisions are to be expected. Hence, the number of collisions will be *at least* as large as the number of bands reported. These two effects will partly cancel

out. We estimate the numbers of collisions given the band count, using the method described in section 3.5.2, and compare with the reported sum of band counts with collisions within and among chromosomes. We quantify correspondence of results by the tail probability of the collision count by Althoff et al. (2007) in our estimated distribution. A small tail probability indicates little correspondence.

		Althoff's results		Our results		
Species	GC content	band count	bands with collisions	expected nr collisions	st.dev.	tail prob
S. cerevisiae	0.38	2	0	0.01	0.12	0.99
		4	0	0.05	0.23	0.95
		3	0	0.03	0.17	0.97
C. elegans	0.36	34	2	3.7	2.1	0.31
		14	1	0.61	0.81	0.44
		17	0	0.90	0.99	0.42
		27	0	2.9	1.6	0.12
A. thaliana	0.36	34	7	3.7	2.1	0.10
		16	2	0.79	0.93	0.19
		10	1	0.32	0.58	0.26
		19	1	1.1	1.1	0.66
D. melanogaster	0.43	27	1	1.5	1.3	0.56
		16	3	0.53	0.75	0.02
		7	0	0.11	0.33	0.90
		29	2	1.8	1.4	0.51
O. sativa	0.44	43	4	3.8	2.1	0.51
		20	3	0.79	0.92	0.05
		24	4	1.1	1.1	0.03
		43	8	3.8	2.1	0.05
M. musculus	0.47	166	66	62.9	10.4	0.39
		134	50	37.7	7.5	0.07
		157	56	54.9	9.5	0.46
		182	86	79.2	12.0	0.29
H. sapiens	0.41	189	87	137.1	19.0	0.002
		134	57	56.0	10.2	0.46
		148	50	71.7	12.0	0.03
		174	86	109.3	16.1	0.07

Table 7.1: Comparison of results on collision of Althoff et al. (2007) and results, based on Gort et al. (2006), for 7 species with different GC contents. A row in the table corresponds to one in-silico AFLP profile. Per profile, the band count and sum of the band counts with collisions within and among chromosomes, reported by Althoff et al. (2007), are shown, together with estimated expectation and standard deviation of the collision count, given the band count, according to Gort et al. (2006). The column labeled "tail prob" gives the tail probability of the result by Althoff et al. (2007) in our estimated distribution.

Results are shown in table 7.1, based on fld's calculated according to the method by Innan et al. (1999) with proper GC-content, as described in section 3.4.1. We conclude that the calculated results on collisions are close to the values reported by Althoff et al. (2007). In one case (first profile from *H. sapiens*), is the reported value of 87 bands with collisions not in accordance with the values we predict. Notice that this is the case with highest number of bands, where we expect the problem of non-comparable results to be largest. Also notice that for *M. musculus* and *H. sapiens* the in-silico profiles contain unrealistically large numbers of bands.

The authors also check the homology of bands between 3 species of *Drosophila*. Their table 2 shows for 8 primer combinations (pc's) the numbers of bands, numbers of pairwise shared bands, and numbers of pairwise homologous bands. From this information we calculate per pc the Dice coefficients for two pairs of profiles, and corresponding Dice coefficients D^{hom} based on homologous bands. We compare D^{hom} with the modified Dice coefficient D^{mod} proposed in section 5.3. Notice that D^{mod} expresses similarity as fraction of homologous *fragments*, whereas D^{hom} is based on *band* counts. Therefore, we do not necessarily expect the same values for D^{hom} and D^{mod}. For a fair comparison, we would need the numbers of homologous fragments, but these are not available. The results are shown in table 7.2 for two pairwise comparisons of *Drosophila* species. D^{mod} gives a slightly stronger correction than D^{hom}.

	melanogaster and simulans			melanogaster and yakuba		
pc	D	D^{hom}	D^{mod}	D	D^{hom}	D^{mod}
caa	0.26	0.22	0.19	0.08	0.04	0.00
cac	0.13	0.13	0.08	0.00	0.00	0.00
cag	0.00	0.00	0.00	0.00	0.00	0.00
cat	0.12	0.08	0.04	0.12	0.08	0.03
tta	0.26	0.15	0.19	0.00	0.00	0.00
ttc	0.12	0.12	0.07	0.06	0.00	0.00
ttg	0.26	0.14	0.13	0.13	0.03	0.00
ttt	0.22	0.07	0.03	0.11	0.00	0.00
average	0.17	0.11	0.09	0.06	0.02	0.004

Table 7.2: Comparison of Dice coefficients for pairs of profiles on *Drosophila* (melanogaster/simulans and melanogaster/yakuba) using 8 pc's (cf. table 2 of Althoff et al. (2007)): D=Dice coefficient, D^{hom} = Dice coefficient based on homologous bands, as reported by Althoff et al. (2007), D^{mod} = Modified Dice coefficient defined in section 5.3.

2. Single nucleotide primer extension
   a) Hansen et al. (1999) report that of 456 investigated bands from 8 pc's on 2 genotypes of *Beta*, 60 bands (13.2%) contained at least two fragments. They do not report the total numbers of bands per pc, nor is it reported whether shared bands among the two genotypes occur only once in the set of 456. The scoring range is also missing. Assuming that all bands are investigated, the average number of bands per profile would equal $456/(2 \times 8) = 28.5$, lower than the 44.3 bands reported as average in the complete study. Assuming a scoring range of 101-550, in-silico Arabidopsis fld F_S (chapter 2), and band count 29, we estimate the number of collisions as 1.9 ± 1.5 (using methods

described in 3.5.2). Assuming no higher order collisions, we estimate that 6.6%±5.2% of the bands contain collisions. The reported 13.2% is well within the 2 σ upper limit. Assuming band count 44, we estimate the number of collisions as 4.4±2.3 or 10%±5.2%. We conclude that our calculated collision counts are comparable to the results by Hansen et al. (1999).

b) O'Hanlon and Peakall (2000a) report that 3 out of 94 bands in profiles on *Carduinae* thistles were amplified by more than one extra primer (collision). A single pc was used, but it remains unclear from how many profiles (genotypes) the 94 bands originate. If there would have been a single profile with 94 bands, our estimated number of collisions (assuming scoring range 101-550, fld from Arabidopsis as before) is 22.8 ± 5.8. The reported 3 bands with collisions is much lower. This could be due to 1) higher order collisions; 2) deficient detection of bands with collisions; 3) multiple profiles, instead of the assumed single profile.

   The authors also report on homoplasy of bands between two genotypes. The amount of relevant information about the AFLP profiles is so limited, however, that we judge it useless to make any predictions.

3. Sequencing of fragments

   a) Rouppe van der Voort et al. (1997), in potato, find homoplasious fragments, even in a selected set of 20 putatively homologous markers. The paper contains incomplete information about the total numbers of bands per profile and the numbers of common bands per pair of profiles, making it impossible to estimate collision counts or counts of homoplasious bands satisfactorily. From their table 3 on sequence comparisons we observe that the size of a fragment differs up to 4.6 base pairs (bp) from the estimated size, based on the position of the band within the lane. We also observe that almost always some internal nucleotides of equally sized fragments from different individuals differ (in the range 0-10 different nucleotides), but these fragments are still called homologous, as they, most likely, originate from the same genomic locus within the different individuals.

   b) Meksem et al. (2001), in soybean, find in a selection of six bands 1-15 different sequences per band (average 6), using 4-30 clones per band. As we understand this, the authors report on average 5 collisions per band. This result contradicts ours, because we would predict the occurrence of 5 collisions to be a highly unlikely event. The high number of collisions may be explained from the fact that the targeted AFLP markers are all linked to one or two loci that confer resistance, in this case resistance to the soybean cyst nematode. Genomic regions conferring resistance are known to harbour repeated DNA sequences. If an AFLP fragment is amplified in this region, neighbouring (almost) identical fragments containing the same restriction sites and selective nucleotides will be amplified as well, resulting in higher than expected collision counts. The authors report that the fragments per band are equally sized (within 1-2 bp).

   c) El-Rabey et al. (2002) sequence 59 bands comigrating for different species of barley. Sequence identity of comigrating bands depends on taxonomic distance between individuals, but also on physical characteristics of the bands (differences in alignment and/or band intensity). Insufficient data are available to do collision or homoplasy calculations. Their finding that more distantly related individuals show more homoplasy (less sequence identity) agrees with our conclusions in chapter 5.

d) Mechanda et al. (2004), in *Echinacea*, also find the relationship between sequence identity and taxonomic distance, by studying one single monomorphic marker of 273 bp (for 79 individuals), and one single polymorphic marker of 159 bp for 48 individuals at 4 taxonomic levels (genus, species, variety, population). For sequencing 1-7 clones per band are taken. Obviously, collision cannot be detected in the case of a single clone. For the bands with at least 2 clones, multiple fragments are often found, but in the case of the monomorphic marker the sequence identities are always higher than 90%, suggesting DNA repeat sequences. For the polymorphic marker, no collision is found in some bands, but in other bands fragments with $\approx 50\%$ sequence identity are reported. For the monomorphic marker band all found fragments have length 273, but for the polymorphic marker with expected size 159 bp, also fragments with lengths 91, 108, and 125 are found. Such large differences in sequence lengths of comigrating bands are not reported by other authors. We could only speculate about the cause of these differences. Maybe hairpin structures of AFLP fragments, or contamination of DNA play a role.

e) Mendelson and Shaw (2005), in crickets, report the occurrence of homoplasy in 1 out 8 sets of comigrating bands without any further details.

f) Ipek et al. (2006), in garlic, study sequence homology of 7 polymorphic AFLP markers in 37 varieties. Two pc's were used, to give 64 and 63 polymorphic markers in the collection of 37 varieties (?, ?). Neither information on the numbers of bands per lane (including monomorphic bands) is given, nor are the numbers of shared bands per pair of varieties mentioned. Therefore, collision and homoplasy calculations are not feasible. For the the 7 markers detailed information about numbers of fragments per band (up to 5), and type of fragment (high sequence or low sequence identity) are given. Fragments with high sequence identity ($> 90\%$) are labeled as homologous. Not all fragments sequenced for one marker have exactly equal lengths. The maximum difference in length is 13.

The general conclusion we draw from these comparisons is that our results are largely, but not always, in accordance with the findings in the literature. For the sequencing studies it is not feasible to estimate collision and homoplasy occurrences due to lack of information. In some of these studies results are obtained, which do not agree fully with our findings. These differences could be due to repeated DNA sequences, or maybe the sequencing of fragments itself introduces extra errors.

For a proper judgement of collision and homoplasy, we recommend that in papers on studies employing AFLP detailed information about scoring range, total numbers of bands per lane, including monomorphic bands, numbers of shared band for pairs of individuals, and band lengths is given. It would be even better to make *all* AFLP information available, e.g. by means of additional web-sources, giving gel-pictures, raw band intensities or peak heights, and interpreted AFLP information, like binary matrices, and codominant scores.

7.3 Relevance for AFLP practice

The problem of homoplasy in AFLP is well known, as may be clear from the compilation of papers in section 1.4, which are revised above. The solutions, suggested by authors, are solutions of restriction: the appliers of AFLP are advised *not* to use AFLP in certain cases, or are advised *not* to score certain bands. Here we cite a number of authors. Althoff et al. (2007) concludes: "AFLP data are best suited for examining phylogeographic patterns within species and among very recently diverged species"; O'Hanlon and Peakall (2000a) concludes: "Studies of phylogeny with AFLPs are therefore only suited to closely related taxa."; Mechanda et al. (2004) even stronger concludes: "Comigrating bands cannot be considered homologous. Thus, the use AFLP band data for comparative studies is appropriate only if results emanating from such analyses are considered as approximations and are interpreted as phenotypic, but not genotypic." Sometimes appliers of AFLP are advised not to use short bands. These advises are reasonable, but could be improved upon. The problem of collision in AFLP is less well known.
We believe the relevance of our work is four-fold:

1. Our results urge appliers of AFLP to become aware of the size of the problems. Appliers may not be aware of the number of collisions that may occur in their profiles. Recognition of the size of the problem will lead to better understanding of the data and its potentially strange behavior. An example would be the strange behavior of some bands in mapping studies. Collision could be the cause of the problem.
2. Refinements in the design of AFLP studies are suggested. If a genotypic interpretation of bands is important, like in QTL studies, it may be better to use highly selective primers, limiting the number of bands per lane. In that case the advise is to go for quality, not for quantity. Our results also allow the applier to pinpoint possibly problematic bands.
3. By modeling the AFLP procedure in a general way, we can quantify the extent of the collision and homoplasy problem, not targeting any special cases. Therefore, we are able to suggest corrections for derived quantities, like the corrected similarity coefficients described in chapter 5.
4. Our work widens the applicability of AFLP. The general advise to use AFLP only for studies of closely related taxa, may be loosened. The problems of collision and homoplasy will always occur, with a smooth transition from small problems in case of AFLP profiles with few bands and closely related taxa, to large problems in case of profiles with many bands and distantly related taxa. The rather artificial dichotomy into situations appropriate for AFLP studies, pretending that problems are non-existing, and inappropriate situations for AFLP studies is suboptimal. Corrections for homoplasy and collisions allow AFLP to be used in a wider range of studies with more reasonable results. This becomes extra relevant at present, where association studies are performed using association panels, consisting of diverse collections of genotypes with little knowledge about their genetic relationships.

7.4 Critical remarks

The core of AFLP is the sampling of DNA fragments from a genome. Next, the fragments need to be identified. To do so, the fragments are separated on an electrophoretic gel or microcapillary system. The DNA molecules have a net negative charge, and migrate within an electric field from the negative to the positive potential through the gel. Longer molecules move slower because they are more easily trapped in the gel. Therefore, the separation of DNA fragments by electrophoresis is mainly by size.

From these facts it seems reasonable to assume that equally long fragments travel equally fast through the lanes of a gel, arriving at the same position within the lanes. In our model of AFLP, we indeed assumed that comigrating fragments are equally sized, hence have equal fragment length probabilities, and that equally sized fragments within a lane appear as a single band. We further assumed that fragments arrive at discrete distances within a lane.

Both assumptions are approximately, but not exactly, right. From the empirical studies it appears that slightly shorter or longer fragments may travel the same distance, may be due to differences in the distribution of the charge. The different studies contradict each other, however, to some extent: Meksem et al. (2001) report that all fragments per band are equally sized, Ipek et al. (2006) find rather small differences in lengths (up to 13 bp), but Mechanda et al. (2004) reports for one of the two studied markers huge differences in lengths (comigrating fragments have lengths in the range 91-161). More study is needed here.

It is not clear how this will influence our results. We could argue that results will remain approximately the same, because some equally long fragments may arrive at a different distance, but some shorter or longer fragments will arrive instead. And hence, the net effect may be approximately nil.

The assumption that fragments arrive at discrete distances corresponding to basepair lengths within a lane is also too simplistic. Maybe with a better scoring algorithm with sub-basepair resolution, part of the homoplasy could be prevented from the start.

Another topic, ignored in our work, is the occurrence of repetitive DNA. If an AFLP fragment is amplified within a repeated stretch of nucleotides, it will be automatically amplified multiple times. We are not inclined to call the occurrence of multiple fragments of the same length within a single lane collision now, because the different fragments are not sampled independently.

7.5 Codominant scoring and collision

So far, the work on collision and homoplasy, as described in chapters 2-5 of the thesis, and the work on codominant scoring, described in chapter 6, are only loosely related. In this section, we sketch how the two topics may be united.

As the band intensity in AFLP is related to the amount of amplified DNA, the intensity is not telling only about the copy number of a DNA fragment, but also about collision. Mixture models with more than 3 components for the band intensity would need to be fitted. An example is shown in figure 7.3. It would be a great challenge to estimate all parameters in this model. The posterior probabilities for an individual to belong to one of the components of the mixture, would partly be determined by collision probabilities, which can be calculated a priori.

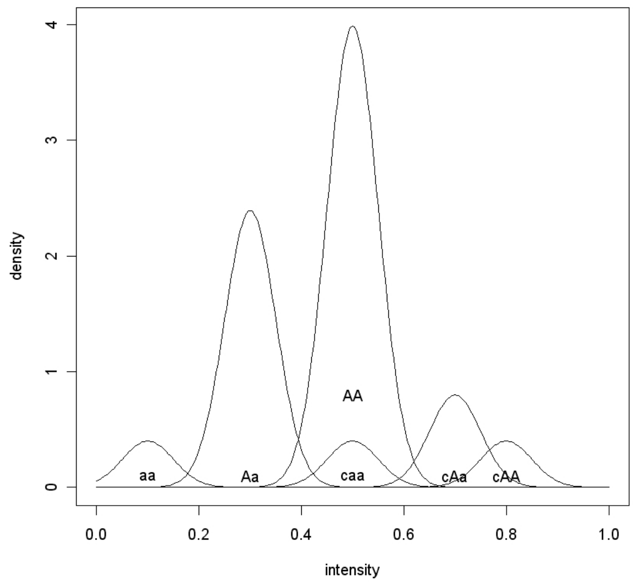


Figure 7.3: Mixture of six normal distributions for band intensities with collisions from a diploid genotype; aa=homozygous absent, Aa = heterozygous, AA = homozygous present, caa = collision of 2 heterozygous fragments, cAa = collision of 1 heterozygous and 1 homozygous fragments, cAA = collision of 2 homozygous fragments.

The collision calculations become more complex now, because we are not dealing with a single individual, but with a collection of related individuals. Below we sketch how collision probabilities in case of a single lane, of two lanes, and more than two lanes can be calculated:

- One lane
  Suppose we have a single AFLP lane with m fragments lengths sampled from a fld F. We are interested in the probability that a collision occurs for a band at position j, i.e. fragment length j, within the lane. Let k_j be the fragment count of fragments with length j, shorter denoted as k. This probability is $P(k > 1|k_1 \geq 1)$, and is easily calculated as $1 - P(k = 1|k \geq 1) = 1 - P(k = 1)/P(k \geq 1) = 1 - \frac{mp_j(1-p_j)^{m-1}}{1-(1-p_j)^m}$, where p_j is the probability of a fragment of length j (cf. formula 4.2). As an example, take $F = F_S$ with scoring range 51-500, and suppose that $m = 80$ fragments were amplified. Suppose that a band at position 3 is present, which corresponds to a relative abundant short

fragment. For this band the collision probability is 0.288. For a band at position 448, which occurs more than 20 times less frequently, the collision probability is only 0.0154.

- Two lanes
  Suppose we have a pair of AFLP lanes, corresponding to two individuals, with equal fragment counts m. The fragment lengths in the two lanes form two related samples from the same fld. We assume that the populations of candidate fragments for the two individuals share a fraction p_c, i.e. the probability of a common fragment is equal to p_c. Since a fraction p_c of the fragments is shared, the presence of a band at position j is likely to have an impact on the collision probability in the first lane. Let k_1 and k_2 be the fragment counts at position j in the first and second lanes. Two probabilities are of interest now:
  1. $P(k_1 > 1|k_1 \geq 1, k_2 \geq 1)$, i.e. the probability of a collision for a band in the first lane, if a band is present in the second lane
  2. $P(k_1 > 1|k_1 \geq 1, k_2 = 0)$, i.e. the probability of a collision in the first lane, without a band in the second lane.

  Using basic calculation rules we can work out these probabilities, using at the right hand side shorthand notation:

$$P(k_1 > 1|k_1 \geq 1, k_2 \geq 1) = 1 - \frac{P(k_1 = 1|k_2 \geq 1)}{P(k_1 \geq 1|k_2 \geq 1)} = 1 - \frac{p_{x|1}}{p_{1|1}}$$

$$P(k_1 > 1|k_1 \geq 1, k_2 = 0) = 1 - \frac{P(k_1 = 1|k_2 = 0)}{P(k_1 \geq 1|k_2 = 0)} = 1 - \frac{p_{x|0}}{p_{1|0}}$$

  We need the conditional probabilities $p_{x|1}, p_{1|1}, p_{x|0}, p_{1|0}$. Using Bayes-rule, and splitting events according to a fragment being sampled from the common part of the two populations of candidate fragments or unique parts of the two populations, we ultimately arrive at:

$$p_{x|1} = \frac{(p_c + p_{\bar{c}j}(1 - p_c))p_{bx}}{p_{b0}}$$

$$p_{1|1} = 1 - \frac{(1 - p_{\bar{c}j})p_{b0}}{1 - p_{b0}}$$

$$p_{x|0} = \frac{(1 - (p_c + p_{\bar{c}j}(1 - p_c)))p_{bx}}{p_{b0}}$$

$$p_{1|0} = 1 - p_{\bar{c}j}$$

  In these formulae we use the binomial probabilities $p_{b0} = P(k_1 = 0) = (1-p_j)^m$, and $p_{bx} = P(k_1 = 1) = p_j(1 - p_j)^{m-1}$; $p_{\bar{c}j} = (1 - \frac{(1-p_c)p_j}{1-p_cp_j})^m$ is the probability that none of the m fragments of lane 2 has length j, given in lane 1 a single fragment of length j not from the common part.
  Here is an example with $p_c = 0.8$ and $m = 80$. For a band at position 3, we find that the probability that a collision occurs given a band in the second lane $P(k_1 > 1|k_1 \geq 1, k_2 \geq 1) = 0.323$, so slightly larger than the unconditional probability 0.288, found earlier. The collision probability, given the absence of

a band in lane 2, is 0.0633. This value is much lower than the unconditional probability.

For less related species, say $p_c = 0.5$, these values become 0.347 and 0.153. For unrelated species with $p_c = 0.0$ the probabilities are equal to the unconditional probability, as the absence or presence of a band in the second lane does not reveal any information about lane 1. For rarely occurring bands, say with length 448, we find for $p_c = 0.8$ collision probabilities 0.0184 and 0.00310.

- Three or more lanes

  Suppose we have three AFLP lanes, corresponding to three individuals with equal fragment counts m. For each lane the fragments are a sample from the fld F, but the three samples are related. We assume that the three populations of candidate fragments share a fraction p_{c123} of the fragments, that populations 1 and 2 share fraction p_{c12}, populations 1 and 3 p_{c13}, and populations 2 and 3 p_{c23}. A fraction p_{c1} of the fragments from population 1 is unique, whereas fraction p_{c2} from population 2, and p_{c3} from population 3 is unique. Notice that $p_{c123}+p_{c12}+p_{c13}+p_{c1} = 1$,$p_{c123}+p_{c12}+p_{c23}+p_{c2} = 1$, and $p_{c123}+p_{c13}+p_{c23}+p_{c3} = 1$.

  As before, focussing on fragments with length j, k_i is the count of fragments (of length j) in lane i. The probabilities that we are interested in, are:

  1. $P(k_1 > 1|k_1 \geq 1, k_2 \geq 1, k_3 \geq 1)$
  2. $P(k_1 > 1|k_1 \geq 1, k_2 \geq 1, k_3 = 0)$
  3. $P(k_1 > 1|k_1 \geq 1, k_2 = 0, k_3 = 0)$

  The question is how the information on absence or presence of bands in other lanes leaks towards the probability of a collision in the first lane. We are able to calculate these probabilities in the general case, using a recursive algorithm. In the formulae for collision probabilities the proportions of common parts of the populations of candidate fragments are assumed to be known, but they are not in practice. We may estimate them by comparing the AFLP lanes, like we did for the pairwise case in chapter 5.

7.6 Future work

Many topics need further attention. We summarize a few here.

- The ideas described in the previous section 7.5 need to be extended to arrive at improved codominant scoring of AFLP based on collision probabilities.
- Combining information from multiple profiles
  Our work so far looked at AFLP profiles from a single primer combination, e.g. to estimate the pairwise genetic similarity. In practice, however, often multiple primer combinations are used, resulting in multiple profiles per individual. It is worthwhile to investigate how the information from multiple primer combinations may be joined to arrive at better estimates.
- Homoplasy corrected versions of AFLP based quantities
  AFLPs are used to estimate e.g. gene diversity or heterozygosity in populations. These quantities will be biased due to collision and homoplasy. It is of great interest to investigate how these quantities may be corrected for collision and homoplasy.
- Other marker systems
  AFLP is just one of many DNA fingerprinting techniques. Other procedures include SSR (microsatellites), RFLP (Restriction Fragment Length Polymorphism), RAPD (Random Amplification of Polymorphic DNA), SNP (Single Nucleotide Polymorphism), SCAR (Sequence Characterized Amplified Region), and DArT (Diversity Arrays Technology). It is worthwhile to investigate how our findings can be applied to other techniques. In AFLP, the basic problem is the incomplete information that we get from electrophoretic gels or microcapillary systems: we only get information about the length of the fragments, not of the sequence identity. This results into the problems of collision and homoplasy. We expect to see the same type of problems in RFLP, RAPD, and possibly SSR's, but not in the other mentioned (more modern) techniques, where electrophoresis does not play a role.
- The effect of unequally sized comigrating fragments
  So far, we assumed that comigrating fragments have equal lengths. How do the estimates of collision and homoplasy change, if we allow comigrating fragments to have lengths deviating from each other?
- The effect of repetitive DNA
  If an AFLP fragment originates from a repeated DNA sequence, multiple copies of the same fragment may be amplified. Especially in codominant scoring this may have consequences.
- The effect of the scoring precision of bands on estimates of collision and homoplasy. If the band position is scored more precisely, i.e. with a resolution higher than 1 bp, how do our estimates of collision and homoplasy change?
- In-silico AFLP
  Over the past years huge databases with genome sequences have been filled, and still are being filled. These resources seem to be underexploited at this moment. Especially the comparison between empirical wet-lab work and in-silico AFLP needs further attention.

7.7 Final remarks

The focus on a molecular marker technique like AFLP, a volatile topic in the dynamic world of molecular biology and genetics, puts our work inevitably at risk of getting early out of date. The cracks in the building called AFLP are visible, as the stabilization of numbers of publications mentioning AFLP seems to indicate. An optimist would say that this grand old lady of DNA fingerprinting is still going strong, but for how long? May be AFLP will be polished up, improved upon, and made ready for another 10 years or so. But gradually and inevitably other marker techniques, maybe SNP, will take over. And also these cannot be kept fresh forever. In the end, what we want to know is the full DNA sequence, not just the bits and pieces sampled by RAPD, RFLP, AFLP, SNP, DArT, or any other DNA fingerprinting technique. The resulting fast, full genomic datasets will be accompanied by fast, powerful methods for extracting relevant information, like the presence of dangerous mutations. These methods will be developed by statisticians, mathematicians, bio-informaticians, or whatever name they will carry by then. We just have to wait longer until full genome scans will be done on a regular basis. You go to the doctor, and while you get undressed, a full genome scan is performed, and a list of all possible genomic risk factors will be spit out by the computer. That will be the future. But at present, our work adds to the already extensive literature on AFLP, aiming at the diagnosis and repair of some of its weaknesses, making the technique fitter and more reliable than before. And, from a general point of view, we see a nice illustration and a fruitful join of quantitative methods and biology, an example of Biometri(c)s at work.

References

Alonso-Blanco, C., Peeters, A. J. M., Koornneef, M., Lister, C., C., D., van den Bosch, N., 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 Journal*, *14*, 259-271.

Althoff, D. M., Gitzendanner, M. A., & Segrave, K. A. (2007). The Utility of Amplified Fragment Length Polymorphisms in Phylogenetics: A Comparison of Homology within and between Genomes. *Systematic Biology*, *56*, 477-484.

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. *Current Opinion in Plant Biology*, *4*, 1-6.

Barow, M., & Meister, A. (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.

Bennett, M. D., Bhandol, P., & Leitch, I. J. (2000). Nuclear DNA amounts in angiosperms and their modern uses— 807 new estimates. *Annals of Botany*, *86*(4), 859–909.

Berloo, R. van, van Heusden, S., Bovy, A., Meijer-Dekens, F., Lindhout, P., & van Eeuwijk, F. (2008). Genetic research in a public-private research consortium: prospects for indirect use of Elite breeding germplasm in academic research. *Euphytica*, *161*, 293-300.

Berloo, R. van, Zhu, A. G., Ursem, R., Verbakel, H., Gort, G., & van Eeuwijk, F. A. (2008). Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes. *Theoretical and Applied Genetics*, *117*, 89-101.

Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms.* New York: Plenum Press.

Blears, M. J., De Grandis, S. A., Lee, H., & Trevors, J. T. (1998). Amplified fragments length polymorphism (AFLP): a review of the procedure and its applications. *Journal of Industrial Microbiology & Biotechnology*, *21*, 99-114.

Bollaerts, K., Eilers, P. H. C., & van Mechelen, I. (2006). Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology*, *59*, 451-469.

Bonin, A., Ehrich, D., & Manel, S. (2007). Statistical analysis of amplified fragment length polymorphism: a toolbox for moleculr ecologists. *Molecular Ecology*, *16*, 3737-3758.

Breyne, P., Dreesen, R., Cannoot, B., Rombaut, D., Vandepoele, K., Rombauts, S., et al. (2003). Quantitatve cDNA-AFLP analysis for genome-wide expression studies. *Molecular genetics and genomics*, *269*(2).

Butler, R. W., & Sutton, R. K. (1998). Saddlepoint approximation for multivariate cumulative distribution functions and probability computations in sampling theory and outlier testing. *Journal of the American Statistical Association*, *93*(442), 596–604.

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.

Castiglioni, P., Ajmone-Marsan, P., van Wijk, R., & Motto, M. (1999). AFLP markers in a molecular linkage map of maize: codominant scoring and linkage group distribution. *Theoretical and Applied Genetics*, *99*, 425-431.

Chakraborty, R. (1993). A class of population genetic questions formulated as the generalized occupancy problem. *Genetics*, *134*, 953–958.

Chen, J., & Li, P. (2009). Hypothesis test for normal mixture models: the EM approach. *Annals of Statistics*, *37*(5A), 2523-2542.

Dasmahapatra, K. K., Hoffman, J. I., & Amos, W. (2009). Pinniped phylogenetic relationships inferred using AFLP markers. *Heredity*, *103*, 168-177.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B - Methodological*, *39*(1), 1–38.

Deniau, A. X., Pieper, B., Ten Bookum, W. M., Lindhout, P., Aarts, M. G. M., & Schat, H. (2006). QTL analysis of cadmium and zinc accumulation in the heavy metal hyperaccumulator *Thlaspi caerulescens*. *Theoretical and Applied Genetics*, *113*, 907-920.

Devos, K. M., Beales, J., Nagamura, Y., & Sasaki, T. (1999). Will colinearity allow gene prediction across the eudicot-monocot divide? *Genome Research*, *9*(7), 825-829.

Dice, L. R. (1945). Measured of the amount of ecological association between species. *Ecology*, *26*, 297-302.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*, *11*, 189-228.

Drossou, A., Katsiotis, A., Leggett, J. M., Loukas, M., & Tsakas, S. (2004). Genome and species relationships in genus *Avena* based on RAPD and AFLP molecular markers. *Theoretical and Applied Genetics*, *109*, 48-54.

Duim, B., Vandamme, P. A. R., Rigter, A., Laevens, S., Dijkstra, J. R., & Wagenaar, J. A. (2001). Differentiation of *Campylobacter* species by AFLP fingerprinting. *Microbiology*, *147*, 2729-2737.

Eck, H. J. van, Rouppe van der Voort, J., Draaistra, J., Zandvoort, P. van, Enckevort, E. van, Segers, B., et al. (1995). The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring. *Molecular Breeding*, *1*, 397-410.

Eeuwijk, F. A. van, & Law, J. R. (2004). Statistical aspects of essential derivation, with illustrations based on lettuce and barley. *Euphytica*, *137*, 129-137.

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–102.

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

Feller, W. (1968). *An introduction to probability theory and its applications volume i.* New York: John Wiley & Sons.

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

Foulley, J. L., van Schriek, M. G. M., Alderson, L., Amigues, Y., Bagga, M., Boscher, M. Y., et al. (2006). Genetic Diversity Analysis Using Lowly Polymorphic Dominant Markers: The Example of AFLP in Pigs. *Journal of Heredity*, *97*, 244-252.

Fraley, C., & Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, *97*(458), 611-631.

Freeling, M. (2001). Grasses as a single genetic system. *Plant Physiology*, *125*, 1191-1197.

Garel, B. (2007). Recent asymptotic results in testing for mixtures. *Computational Statistics & Data Analysis*, *51*, 5295-5304.

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R. L., Dunn, M., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp japonica). *Science*, *296*(5565), 92-100.

Gort, G., Koopman, W. J. M., & Stein, A. (2006). Fragment length distributions and collision probabilities for AFLP markers. *Biometrics*, *62*, 1107-1115.

Gort, G., Koopman, W. J. M., Stein, A., & van Eeuwijk, F. A. (2008). Collision Probabilities for AFLP Bands, With an Application to Simple Measures of Genetic Similarity. *Journal of Agricultural, Biological, and Environmental Statistics*, *13*(2), 177-198.

Hansen, M., Kraft, T., Christiansson, M., & Nilsson, N. O. (1999). Evaluation of AFLP in Beta. *Theoretical and Applied Genetics*, *98*, 845–852.

Henze, N. (1998). A Poisson limit law for a generalized birthday problem. *Statistics & Probability Letters*, *39*, 333–336.

Holland, B. R., Clarke, A. C., & Meudt, H. M. (2008). Optimizing Automated AFLP Scoring Parameters to Improve Phylogenetic Resolution. *Systematic Biology*, *57*, 347-366.

Holst, L. (1995). The general birthday problem. *Random Structures and Algorithms*, *6*(2 and 3), 201–208.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299–314.

Imazio, S., Labra, M., Grassi, F., Winfield, M., Bardini, M., & Scienza, A. (2002). Molecular tools for clone identification: the case of the grapevine cultivar 'Traminer'. *Plant Breeding*, *121*, 531–535.

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

Ipek, M., Ipek, A., & Simon, P. W. (2006). Sequence homology of polymorphic AFLP markers in garlic (*Allium sativum* L.). *Genome*, *49*, 1246-1255.

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. *Theoretical and Applied Genetics*, *114*, 421-428.

Jansen, R. C. (1993). Maximum Likelihood in a Generalized Linear Finite Mixture Model by Using the EM Algorithm. *Biometrics*, *49*, 227-231.

Jansen, R. C. (1994). Maximum Likelihood in a finite mixture model by exploiting the GLM facilities of Genstat. *Genstat Newsletter*, *30*, 25-27.

Jansen, R. C., Geerlings, H., van Oeveren, A. J., & van Schaik, R. R. (2001). A comment on Codominant Scoring of AFLP markers. *Genetics*, *158*, 925-926.

Jeuken, M. J. W., & Lindhout, P. (2004). The development of lettuce backcross inbred lines (BILs) for exploitation of the *Lactuca saligna* (wild lettuce) germplasm. *Theoretical and Applied Genetics*, *109*, 394-401.

Jeuken, M. R., Van Wijk, R., Peleman, J., & Lindhout, P. (2001). An integrated interspecific AFLP map of lettuce (*Lactuca*) based on two *L-sativa* × *L-saligna* F-2 populations. *Theoretical and Applied Genetics*, *103*, 638-647.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). *Univariate discrete distributions.* New York: John Wiley & Sons.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In H. Munro (Ed.), *Mammalian protein metabolism* (p. 21-132). New York: Academic Press.

Karp, A., Seberg, O., & Buiatti, M. (1996). Molecular techniques in the assessment of botanical diversity. *Annals of Botany*, *78*, 143-149.

Kathman, S. J., & Terrell, G. R. (2003). Poisson approximation by constrained exponential tilting. *Statistics & Probability Letters*, *61*, 83–89.

Keygene products B.V. (2004). AFLP-Quantar(r)Pro 1.0 - Part I - User Guide [Computer software manual]. Wageningen, The Netherlands.

Koopman, W. J. M. (2002). *Zooming in on the lettuce genome.* Unpublished doctoral dissertation, Wageningen University. (Chapter 6, Evolution of DNA content and base composition in *Lactuca* (Asteraceae) and related genera)

Koopman, W. J. M., & 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, W. J. M., Zevenbergen, M. J., & Van den Berg, R. G. (2001). Species relationships in *Lactuca* s.l. (Lactuceae, Asteraceae) inferred from AFLP fingerprints. *American Journal of Botany*, *88*(10), 1881-1887.

Kosman, E., & Leonard, K. J. (2005). Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Molecular Ecology*, *14*, 415–424.

Kraus, S. L. (2000). Accurate gene diversity estimates from amplified fragment length polymorphism (AFLP) markers. *Molecular Ecology*, *9*, 1241-1245.

Levin, B. (1981). A representation for multinomial cumulative distribution functions. *The Annals of Statistics*, *9*(5), 1123–1126.

Lindsay, B. G. (1995). *Mixture models: theory, geometry and applications.* Hayward: Institute of Mathematical Statistics.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology.* London: Chapman and Hall.

Marie, D., & Brown, S. C. (1993). A cytometric exercise in plant DNA histograms, with 2C-values for 70 species. *Biology of the Cell*, *78*(1-2), 41–51.

Matassi, G., Montero, L. M., Salinas, J., & Bernardi, G. (1989). The isochore organizaion and the compositional distribution of homologous coding sequences in the nuclear genome of plants. *Nucleic Acids Research*, *17*(13), 5273-5290.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Second ed.).

Chapman & Hall.

McGregor, C. E., van Treuren, R., Hoekstra, R., & van Hintum, T. J. L. (2002). Analysis of the wild potato germplasm of the series Acaulia with AFLPs: implications for ex situ. *Theoretical and Applied Genetics*, *104*, 146-156.

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York: John Wiley & sons.

McLachlan, G. J., Ng, S.-K., & Bean, R. (2006). Robust Cluster Analysis via Mixture Models. *Austrian Journal of Statistics*, *35*(2-3), 157-174.

Mebrate, S. A., Dehne, H. W., Pillen, K., & Oerke, E. C. (2006). Molecular diversity in *Puccinia triticina* isolates from Ethiopia and Germany. *Journal of Phytopathology*, *154*(11-12), 701-710.

Mechanda, S. M., Baum, B. R., Johnson, D. A., & Arnason, J. T. (2004). Sequence assessment of comigrating AFLP bands in *Echinacea* - implications for comparative biological studies. *Genome*, *47*, 15-25.

Meksem, K., Ruben, E., Hyten, D., Triwitayakorn, K., & Lightfoot, D. A. (2001). Conversion of AFLP bands into high-throughput dna markers. *Molecular Genetics and Genomics*, *265*, 207-214.

Mendelson, T. C., & Shaw, K. L. (2005). Use of AFLP Markers in Surveys of Arthropod Diversity. *Methods in Enzymology*, *395*, 161-177.

Meudt, H. M., & Clarke, A. C. (2007). Almost Forgotten or Latest Practice? AFLP applications, analyses and advances. *Trends in Plant Science*, *12*, 106-117.

Montero, L. M., Salinas, J., Matassi, G., & Bernardi, G. (1990). Gene distribution and isochore organization in the nuclear genome of plants. *Nucleid Acids Research*, *18*(7), 1859-1867.

Mueller, U. G., & LaReesa Wolfenbarger, L. (1999). AFLP genotyping and fingerprinting. *Trends in Ecology and Evolution*, *14*, 389-394.

Munford, A. G. (1977). A note on the uniformity assumption in the birthday problem. *The American Statistician*, *31*, 119.

Nagl, W., & Stein, B. (1989). DNA characterization in host-specific *Viscum album* subspecies (Viscaceae). *Plant Systematics and Evolution*, *166*, 243-248.

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*, 5269-5273.

Nussinov, R. (1981). Nearest neighbor nucleotide patterns - structural and biological implications. *Journal of Biological Chemistry*, *256*, 8458-8462.

Nussinov, R. (1991). Compositional variations in DNA sequences. *Computer Applications in the Biosciences*, *7*, 287-293.

O'Hanlon, P. C., & Peakall, R. (2000a). A simple method for the detection of size homoplasy among amplified fragment length polymorphism fragments. *Molecular Ecology*, *9*, 815-816.

O'Hanlon, P. C., & Peakall, R. (2000b). A simple method for the detection of size homoplasy among amplified fragment length polymorphism fragments. *Molecular Ecology*, *9*, 815-816.

Peters, J. L., Constandt, H., Neyt, P., Cnop, G., Zethof, J., Zabeau, M., et al. (2001). A physical amplified fragment-length polymorphism map of *Arabidopsis. Plant Physiology*, *127*, 1579-1589.

Piepho, H. P., & Koch, G. (2000). Codominant analysis of banding data from a dominant marker system by normal mixtures. *Genetics*, *155*, 1459-1468.

Prochazka, M., Walder, K., & Xia, J. (2001). AFLP fingerprinting of the human genome. *Human Genetics*, *108*, 59-65.

R Development Core Team. (2005). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

Reamon-Büttner, S. M., Schondelmaier, J., & Jung, C. (1998). AFLP markers tightly linked to the sex locus in *Asparagus officinalis* L. *Molecular Breeding*, *4*, 91-98.

Reif, J. L., Melchinger, A. E., & Frisch, M. (2005). Genetical and Mathematical Properties of Similarity and Dissimilarity Coefficients Applied in Plant Breeding and Seed Bank Management. *Crop Science*, *45*, 1-7.

Rieseberg, L. H. (1996). Homology among RAPD fragments in interspecific comparisons. *Molecular Ecology*, *5*, 99-105.

Robinson, J. P., & Harris, S. A. (1999). Amplified fragment length polymorphisms and microsatellites: A phylogenetic perspective. In E. M. Gillet (Ed.), *Which DNA marker for which purpose?* (e-book http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm)

Rohlf, F. J. (1993). *Ntsys-pc, numerical taxonomy and multivariate analysis system.* Setauket, NY: Exeter Software.

Rouppe van der Voort, J. N. A., van Zandvoort, P., van Eck, H. J., Folkertsma, R. T., Hutten, R. C. B., Draaistra, J., et al. (1997). Use of allele specificity of comigrating AFLP markers to align genetic maps from different potato genotypes. *Molecular & General Genetics*, *255*, 438-447.

Salinas, J., Matassi, G., Montera, L. M., & Bernardi, G. (1988). Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. *Nucleic Acids Research*, *16*(10), 4269-4285.

Sandell, D. (1991). Computing probabilities in a generalized birthday problem. *The Mathematical Scientist*, *16*, 78-82.

Sasaki, T., Matsumoto, T., Yamamoto, K., Sakata, K., Baba, T., Katayose, Y., et al. (2002). The genome sequence and structure of rice chromosome 1. *Nature*, *420*(6913), 312-316.

Silvey, S. D. (1975). *Statistical Inference.* London: Chapman and Hall.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy.* San Francisco, CA: Freeman.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy.* San Francisco/London: W.H. Freeman.

Tams, S. H., Melchinger, A. E., & Bauer, E. (2005). Genetic similarity among European winter triticale elite germplasms assessed with AFLP and comparisons with SSR and pedigree data. *Plant Breeding*, *124*(2), 154-160.

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. *Molecular Ecology*, *11*, 139–151.

Venzon, D. J., & Moolgavkar, S. H. (1988). A Method for Computing Profile-Likelihood-Based Confidence Intervals. *Applied Statistics*, *37*, 87-94.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Vandelee, T., Hornes, M., et al. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acid Research*, *23*(21), 4407–4414.

Vuylsteke, M. (2007). AFLP technology for DNA fingerprinting. *Nature Protocols*, *2*, 1387-1398.

Wenzl, P., Carling, J., Kudrna, D., Jaccoud, D., Huttner, E., Kleinhofs, A., et al. (2004). Diversity Arrays Technology (DArT) for whole-genome profiling of barley. *Proceedings of the National Academy of Sciences USA*, *101*, 9915-9920.

Wong, A., Forbes, M. R., & Smith, M. L. (2007). Characterization of AFLP markers in damselflies: prevalence of codominant markers and implications for population genetic applications. *Genome*, *44*, 677-684.

Yu, J., Hu, S. N., Wang, J., Wong, G. K. S., Li, S. G., Liu, B., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp indica). *Science*, *296*(5565), 79-92.

Zhanjiang, L. (2007). Amplified Fragment Length Polymorphism (AFLP). In L. Zhangjiang (Ed.), *Aquaculture genome technologies* (p. 29-42). Ames, Iowa: Blackwell Publishing.

Zhong, D., Menge, D. M., Temu, E. A., 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.

Summary

AFLP is a DNA fingerprinting technique, that is used in a wide variety of genetic applications. An AFLP fingerprint (or profile) of an individual, be it a plant, bacterium, yeast, animal, or human being, consists of bands, visible on different positions within a lane of an electrophoretic gel or microcapillary system. The bands represent DNA fragments sampled from the genome. Profiles are usually interpreted as binary band presence - absence patterns, making AFLP bands dominant markers: they do not distinguish between homozygous (AA) and heterozygous (Aa) situations. It is assumed that equally long fragments travel equally far (comigrate) through the lanes of a gel by electrophoresis. Therefore, the position of a band within a lane indicates the length of the underlying DNA fragment.

By comparing the profiles of two individuals, the pairwise genetic similarity between two individuals can be determined, which is one of the many applications of AFLP. Usually some of the bands of the pair of profiles are shared, whereas others are unique. Commonly used similarity coefficients are Dice and Jaccard similarities, in which the fraction of shared bands is calculated in different ways.

Comigrating bands occur if in two individuals an identical DNA fragment originating from the same genomic locus is amplified. It is also possible, however, that two equally sized fragments, but of different nucleotide composition and of different genomic origin, were amplified and comigrate. This type of band sharing by chance is called *homoplasy.* Homoplasy is undesirable: we see two corresponding bands, but the correspondence is false. Homoplasy will e.g. bias similarity coefficients.

Another type of homoplasy is the comigration of equally sized fragments of different nucleotide composition within a lane, i.e. for a single individual. To distinguish this type of homoplasy from the first, we call it *collision.* Like homoplasy, collision is undesirable: we interpret a single band as a single fragment, but two or more fragments are hidden within the band.

The main topic in this thesis is the study of collision and homoplasy in AFLP. We answer questions like: How often do they occur? What are the main determinants? What are the possible consequences, and how can we correct for them? To answer these questions we model the AFLP procedure. The first step of the AFLP procedure is the creation of a population of candidate fragments, by cutting the genome into fragments by restriction enzymes. Depending on the genome size, this population may contain millions of fragments. The frequency distribution of the lengths of the candidate fragments is called the fragment length distribution (fld). Only fragments with lengths within a scoring range (e.g. 50-600) are scored. The next step is the (random) sampling of fragment lengths from the fld, using primers with selective nucleotides. The last step is the binary scoring of the bands,

only indicating whether *at least* one fragment of a given length is present.

The fld plays a central role in the study of homoplasy and collision. We study it from different perspectives: in chapter 2 from theoretical considerations (cf. Innan et al., 1999) and by an in-silico approach, and in chapter 3 by estimating it from the AFLP profile itself, using a monotonic smoother and generalized linear models. For the in-silico approach the AFLP procedure was mimicked on the computer and applied to the available genome sequences of *Arabidopsis thaliana* and *Oryza sativa*. The fld is highly asymmetric, with shorter fragments much more abundant than long fragments.

The research in this thesis starts off with a phylogenetic study in lettuce, where the question is raised whether similarity between species calculated from binary AFLP profiles can be due to chance alone. In chapter 2 we answer this question using a Monte Carlo approach. We simulate the distribution of similarity coefficients for unrelated individuals, that is, assuming that all band sharing is caused by chance. We find that chance similarity can be extensive, mainly depending on the number of bands in the lanes. For instance, for two lanes with 120 bands each the average Dice coefficient is 0.4. Critical regions to test the null hypothesis of unrelatedness are derived. Also, weighted similarity coefficients are suggested.

Based on findings of chapter 2, a theoretical study on collisions is done, described in chapter 3. The collision problem is analogous to the birthday problem, telling that only 23 persons are needed to have a probability of a shared birthday of more than $^1/_2$. It is a generalized birthday problem, because, unlike the birthday distribution, the fld is not uniform. For a typical plant genome, an AFLP with 19 bands is likely to contain a first collision. A profile with 100 bands may contain 25 (± 6) collisions. The distribution of the total collision count in a profile is calculated for three situations: 1) given the fragment count, 2) given the band count, and 3) given the band lengths (in chapter 4). For known fragment count, the distribution is a generalized occupancy distribution, approximated by a binomial distribution. The probability of no collision is a multinomial tail probability, calculated by a saddlepoint approximation. Larger collision counts are found for profiles with more bands, more skewed fld, and smaller scoring range.

Chapter 4 describes a continued study on collision, now focusing on the collision probability for individual bands. We demonstrate how the probability of no collison for an individual band is calculated for the above mentioned three situations. Since short fragments occur more often, short bands are more likely to contain collisions. For a typical plant genome and AFLP procedure, the collision probability for the shortest band is 25 times larger than for the longest. The findings are summarized in a list of recommendations for AFLP practice. We show how collision calculations can be used to get modified Dice and Jaccard similarity coefficients, corrected for collision an homoplasy.

In chapter 5 the topic of homoplasy corrected estimation of pairwise genetic similarity is studied further. Estimators Dice (D) and Jaccard overestimate genetic similarity, due to homoplasy. The bias of D increases with larger numbers of bands, and lower genetic similarity. We propose two estimators of genetic similarity, which correct for homoplasy and collision. Properties of the estimators are studied by simulation and bootstrapping. The estimators are nearly unbiased, and have for most practical cases smaller standard error than D. The relationship be-

tween fragment counts and precision is studied using simulation. The usual range of band counts (50-100) appears nearly optimal.
Chapter 6 describes a study on the codominant scoring of AFLP markers in association panels. In codominant scoring the intensity of a band is classified into one of three groups (AA, Aa, aa), by fitting a normal mixture model. Association panels are collections of individuals without prior information on genotype probabilities. We study features to improve or stabilize the unmixing of the band intensities, and diagnostics for data quality. Our approach provides posterior genotype probabilities for marker loci, that can form the basis for association mapping. Software has been developed in R, containing the models for normal mixtures with facilitating features, and visualizations. The methods are applied to an association panel in tomato (which is part of a larger study within the Dutch Center for BioSystem Genomics).
The connection between chapters $2-5$ on collision and homoplasy, and chapter 6 on codominant scoring is reinforced in the discussion chapter 7, hinting on how collision probabilities may be used in mixture models. It is also described how collision probabilities depend on information from other lanes. Examples of AFLP in lettuce and tomato serve as illustrations throughout the manuscript.

Samenvatting

AFLP is een DNA fingerprinting techniek, die veel toepassingen kent. Een AFLP fingerprint (ook wel: profiel) van een individu, zij het een plant, bacterie, gist, dier, of mens, bestaat uit bandjes, die zichtbaar zijn op verschillende posities in een laan van een electroforetische gel of microcapillair systeem. De bandjes stellen DNA fragmenten voor, verkregen door het genoom te bemonsteren. Het is gebruikelijk dat profielen geïnterpreteerd worden als binaire aan- en afwezigheidspatronen van bandjes. In dat geval zijn AFLP-bandjes dominante markers: ze maken geen onderscheid tussen homozygote (AA) en heterozygote (Aa) situaties. We veronderstellen dat bij electroforese even lange fragmenten even snel door de lanen van de gel reizen (comigreren). Daardoor is de positie van een bandje binnen een laan een indicatie voor de lengte van het achterliggende DNA fragment.
Door de profielen van twee individuen te vergelijken, kan de paarsgewijze genetische similariteit tussen twee individuen worden bepaald. Dit is één van de vele toepassingen van AFLP. Gewoonlijk is een aantal bandjes van een paar profielen gemeenschappelijk, terwijl de overige uniek zijn. Gebruikelijke similariteitscoëfficiënten zijn Dice en Jaccard similariteiten, die op verschillende manieren de fractie gemeenschappelijke bandjes berekenen.
Comigrerende bandjes treden op als bij twee individuen een identiek DNA fragment van dezelfde positie op het genoom is geamplificeerd. Het is echter ook mogelijk, dat twee even grote fragmenten geamplificeerd zijn en comigreren, terwijl ze een verschillende nucleotidesamenstelling hebben en afkomstig zijn van verschillende posities op het genoom. Het verschijnsel dat twee verschillende banden comigreren en als een enkele band worden geïnterpreteerd, heet *homoplasie.* Homoplasie is ongewenst: we zien twee corresponderende banden, maar de correspondentie is schijn. Door homoplasie overschatten similariteitscoëfficiënten de werkelijke similariteit.
Een ander type homoplasie is de comigratie van even lange fragmenten van verschillende nucleotidesamenstelling binnen één laan, dat wil zeggen voor een enkel individu. Om dit type homoplasie te onderscheiden van het eerste type, noemen we het *collisie.* Net zoals homoplasie, is collisie ongewenst: we interpreteren een bandje als een enkel fragment, maar twee of meer fragmenten zijn verscholen in het bandje.
Het hoofdonderwerp van dit proefschrift is de studie van collisie en homoplasie in AFLP. We beantwoorden vragen zoals: Hoe vaak treden collisie en homoplasie op? Wat zijn de belangrijkste determinanten? Wat zijn de mogelijke gevolgen, en hoe kunnen we er voor corrigeren? Om deze vragen te beantwoorden, hebben we AFLP gemodelleerd. In de eerste stap van de AFLP procedure wordt het genoom in fragmenten geknipt met behulp van restrictie-enzymen, waardoor een populatie

van kandidaatfragmenten ontstaat. Afhankelijk van de grootte van het genoom, kunnen zo tot miljoenen fragmenten gevormd worden. De frequentieverdeling van de lengte van de kandidaatfragmenten heet de fragment lengte verdeling (fld). Alleen fragmenten met lengtes in een scoringsbereik (bijvoorbeeld 50-600) worden op de gel gescoord. Vervolgens wordt een (aselecte) steekproef van fragmentlengtes uit de fld getrokken met behulp van primers met selectieve nucleotiden. In de laatste stap worden de bandjes binair gescoord. De binaire score geeft aan of er *minstens één* fragment van een bepaalde lengte aanwezig is.

De fld speelt een centrale rol in de studie van collisie en homoplasie. We hebben de fld vanuit verschillende perspectieven bestudeerd: in hoofdstuk 2 vanuit theoretische overwegingen (zie Innan et al., 1999) en door middel van een in-silico benadering, en in hoofdstuk 3 door de fld rechtstreeks te schatten vanuit het AFLP profiel zelf, met behulp van een monotone gladde functie en gegeneralizeerde lineaire modellen. Voor de in-silico aanpak hebben we de AFLP procedure op de computer nagespeeld, en toegepast op de beschikbare genoomsequenties van *Arabidopsis thaliana* en *Oryza sativa*. De fld is asymmetrisch, waarbij korte fragmenten veel vaker voorkomen dan lange.

Het onderzoek in dit proefschrift begint met een phylogenetische studie van sla. In dit onderzoek wordt de vraag gesteld of de berekende similariteiten tussen soorten op basis van binaire AFLP profielen louter door toeval verklaard kunnen worden. In hoofdstuk 2 beantwoorden we deze vraag door middel van een Monte Carlo aanpak. We simuleren de kansverdeling van enkele similariteitscoëfficiënten voor niet-gerelateerde individuen, dat wil zeggen, veronderstellend dat iedere comigratie van banden op toeval berust. We vinden dat kanssimilariteit groot kan zijn, en vooral afhangt van het aantal bandjes in de lanen. Bijvoorbeeld de gemiddelde Dice coëfficiënt bij twee lanen, ieder met 120 bandjes, is 0.4. Kritieke waarden om de nulhypothese van ongerelateerdheid te toetsen zijn afgeleid. Tevens zijn gewogen similariteitscoëfficiënten geïntroduceerd.

Op basis van de bevindingen van hoofdstuk 2, is een theoretische studie uitgevoerd, zoals beschreven in hoofdstuk 3. Het collisieprobleem is analoog aan het verjaardagsprobleem, waarvan de oplossing luidt dat er slechts 23 personen nodig zijn, zodat de kans, dat er minstens twee mensen zijn met dezelfde verjaardag, groter is dan $^1/_2$. Het probleem bij collisie is een gegeneralizeerd verjaardagsprobleem, want, in tegenstelling tot de kansverdeling van verjaardagen, is de fld niet uniform. Een AFLP met slechts 19 bandjes heeft, bij een typisch plantengenoom, een kans op minstens één collisie groter dan $^1/_2$. Een profiel met 100 bandjes kan 25 (±6) collisies bevatten. De kansverdeling van het totaal aantal collisies in een profiel is bepaald voor drie situaties: 1) gegeven het totale aantal fragmenten, 2) gegeven het totale aantal bandjes, en 3) gegeven de bandlengtes (in hoofdstuk 4). Voor bekend fragmentaantal, is deze kansverdeling een gegeneralizeerde occupancy-verdeling, benaderd met een binomiale verdeling. De kans dat in een profiel geen collisie optreedt, is een multinomiale staartkans, berekend met behulp van een zadelpuntbenadering. Een groter aantal collisies treedt op bij profielen met meer bandjes, schevere fld, en kleiner scoringsbereik.

In hoofdstuk 4 is de studie van collisie vervolgd, maar nu met nadruk op de collisiekansen voor individuele bandjes. We laten zien hoe de kans dat geen collisie optreedt voor een individueel bandje kan worden berekend voor de drie bovengenoemde situaties. Omdat korte fragmenten vaker voorkomen, hebben ze een grotere kans op collisie. De collisiekans voor de kortste band is 25 keer groter dan voor de langste band, bij een typisch plantengenoom en standaard AFLP procedure. We hebben onze bevindingen samengevat in een aantal aanbevelingen voor de praktijk. We laten ook zien hoe collisieberekeningen gebruikt kunnen worden om gemodificeerde Dice en Jaccard similariteiten te verkrijgen, die corrigeren voor collisie en homoplasie.

In hoofdstuk 5 wordt het homoplasie-gecorrigeerd schatten van paarsgewijze genetische similariteit verder uitgediept. De Dice (D) en Jaccard similariteiten overschatten genetische similariteit ten gevolge van homoplasie. De onzuiverheid van D neemt toe met grotere aantallen bandjes, en met lagere genetische similariteit. We introduceren twee schatters van genetische similariteit, die corrigeren voor homoplasie en collisie. Enkele eigenschappen van deze schatters zijn bestudeerd door middel van simulatie en de bootstrap. De schatters zijn nagenoeg zuiver, en hebben in de meeste praktische gevallen een kleinere standaardfout dan D. Het verband tussen aantallen fragmenten en precisie is onderzocht via simulatie. Het blijkt dat het gebruikelijke aantal bandjes (50-100) bijna optimaal is.

Hoofdstuk 6 beschrijft een studie over het codominant scoren van AFLP markers in associatiepanelen. Bij codominant scoren wordt de intensiteit van een bandje geclassificeerd in één van de drie genotype klassen AA, Aa, of aa, door het aanpassen van een normaal mengselmodel. We bedoelen met associatiepanelen groepen individuen zonder *a priori* informatie over genotypekansen. We introduceren mogelijkheden om het ontmengen van de intensiteiten te verbeteren of te stabiliseren, en diagnostische grootheden voor datakwaliteit met betrekking tot het codominant scoren. Onze aanpak levert posterior genotypekansen voor marker loci op, die als basis kunnen dienen voor verdere associatiestudie. We hebben software in R ontwikkeld, die de modellen voor normale mengsels met faciliterende opties en visualisaties bevat. De methoden zijn toegepast op een associatiepaneel van tomaten (dat onderdeel is van een grotere studie binnen het Nederlandse Center for BioSystems Genomics).

De samenhang tussen hoofdstukken $2-5$ over collisie en homoplasie, en hoofdstuk 6 over codominant scoren wordt verder uitgewerkt in hoofdstuk 7. Hierin wordt aangegeven hoe collisiekansen benut kunnen worden in mengselmodellen. Tevens wordt besproken hoe collisiekansen afhangen van informatie uit andere lanen. AFLP voorbeelden van sla en tomaat dienen op vele plaatsen als illustratie in het manuscript.

Curriculum vitae

Gerrit Gort was born on April 14th, 1961 in Amsterdam, The Netherlands, as second child in a family of four children. He went to the "Pro Rege" primary school in Amsterdam, and at the secondary school "Christelijke Scholengemeenschap Pascal" he passed the final examination for Gymnasium-β in 1979. At that moment in life he thought that being a medical doctor would be great. Hence he started in 1979 a study Medicine at the Free University of Amsterdam, where he did his kandidaatsexamen in 1982. Two years after the start of the medical study, he was lured back into his old love Mathematics. A study Mathematics with specialization Applied Statistics was commenced in 1981, and finalized in 1987. Worth mentioning are his traineeship at the Center of Quantitative Methods of Philips in Eindhoven, and a first degree qualification to teach Mathematics at secondary schools. After military service, he started his professional career as a statistician at the Group of Biostatistics, Epidemiology and Theory of Medicine of the Medical Faculty at the Free University in Amsterdam. In 1990 a position as university lecturer at Wageningen University was obtained, that lasts until now. Until the year 2000, he was mainly active in statistical consultation at Wageningen University, with teaching at second place. Since 2000 there has been a shift into more research oriented work. The project, resulting in the present PhD-thesis, was commenced on Jan 1, 2001. The author is member of the Vereniging voor Statistiek and of the Netherlands Region of the International Biometric Society, and was active in the societal life of the latter, resulting in the organization of the International Biometric Conference Amsterdam in 1996. Gerrit Gort has a partner and two daughters, and lives in Utrecht.

PE&RC PhD Education Certificate

With the educational activities listed below the PhD candidate has complied with the educational requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of Literature (5.6 ECTS)
- Review AFLP literature with focus on homoplasy

Writing of Project Proposal (7 ECTS)
- Statistical methods for genebanks
- Statistical properties of AFLP

Laboratory Training and Working Visits (4.3 ECTS)
- LD mapping in populations of lettuce; statistical properties of AFLP (WEHI, Melbourne)

Post-Graduate Courses (4.0 ECTS)
- Advanced Linkage Analysis (LUC, 2002)
- VGAM (WUR, 2003)
- Association-based methods of gene mapping (LUC, 2003)
- Smoothing course (P. Eilers, 2004)

Deficiency, Refresh, Brush-up Courses (1.5 ECTS)
- Preconference course R (Dublin, IBS, 2008)
- Preconference course Bayesian Statistics (Montreal, IBS, 2006)
- Preconference course Smoothing (Ghent, IBS-ANed, 2009)

Competence Strengthening / Skills Courses (1.4 ECTS)
- Time Management (2003)

Discussion Groups / Local Seminars and Other Scientific Meetings (7 ECTS)
- Biometrics colloquia (2002-2009)

PE&RC Annual Meetings, Seminars and the PE&RC Weekend (2.3 ECTS)
- PE&RC annual meeting(2004)
- Support PhD students following PE&RC (stats) courses

International Symposia, Workshops and Conferences (7 ECTS)
- IBC (Dublin, 2008)
- IBC (Montreal, 2006)
- RSS (Diepenbeek, 2003)
- Eucarpia Meeting (Dundee, 2009)

Courses in Which the PhD Candidate Has Worked as a Teacher (2 students; 30 days)

-

Dankwoord

Dit proefschrift is tot stand gekomen dankzij directe, en vaak indirecte, bijdragen van velen. Werkend bij een groep als Biometris is het gevaar van versnippering van tijd groot. Vele onderwerpen vragen de aandacht. Op de eerste plaats staat het onderwijs in de statistiek aan BSc-, MSc-, en ook PhD-studenten. Onderwijs is, naar mijn idee, dé reden van bestaan van de leerstoelgroep Statistiek aan Wageningen Universiteit. Naast het voorbereiden en geven van onderwijs, vraagt vernieuwing van onderwijs continu de aandacht. Op de tweede plaats staat de consultatie. Vele onderzoekers kampen met vragen over het toepassen van de statistiek in hun eigen situatie, en komen bij Biometris langs om die vragen te bespreken. Op de laatste plaats, zo lijkt het vaak, staat het eigen onderzoek, hoewel daarin in de laatste jaren verandering is gekomen. Toch is deze laatste post nodig om tot een proefschrift te komen. Ik wil Johan Grasman en Fred van Eeuwijk, als toenmalig en huidig hoofd van Biometris, bedanken dat me de ruimte is gegeven om het onderzoek te verrichten en af te ronden. Alfred Stein wil ik bedanken voor zijn energieke, aanmoedigende rol gedurende de afgelopen jaren. Daarbij aansluitend is een dankwoord op zijn plaats gericht aan alle statistiek collega's, omdat zij me in de rustige onderwijsperiodes de tijd gaven om aan het onderzoek te werken. Daarbij heb ik wel eens wat licht honende opmerkingen moeten verdragen over de toch wel lange duur van het onderzoek.

Wim Koopman wil ik bedanken voor de samenwerking. Wim, het is het onderwerp van jouw eigen promotie-onderzoek geweest, dat uiteindelijk geleid heeft tot het huidige proefschrift.

Tot slot wil ik de familie bedanken: Mariëtte, Laurien en Sophie, bedankt voor alle ondersteuning gedurende de afgelopen jaren.