Theor Appl Genet (2010) 121:337–351
DOI 10.1007/s00122-010-1313-x

ORIGINAL PAPER

Codominant scoring of AFLP in association panels

Gerrit Gort · Fred A. van Eeuwijk

Received: 27 November 2009 / Accepted: 1 March 2010 / Published online: 17 March 2010

Abstract 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 band intensities, illustrating and optimizing existing methodology, which employs 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-component 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 1,175 polymorphic markers on 94 tomato hybrids, as part of a larger study within the Dutch Centre for BioSystems Genomics.

Communicated by M. Sillanpaa.

G. Gort (✉) · F. A. van Eeuwijk
Biometris, P.O. Box 100, 6700 AC Wageningen,
The Netherlands
e-mail: gerrit.gort@wur.nl

F. A. van Eeuwijk
Centre for BioSystems Genomics, P.O. Box 98,
6700 AB Wageningen, The Netherlands

Introduction

Amplified fragment length polymorphism (AFLP) (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, i.e., 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, 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 et al. (1999), Reamon-Büttner et al. (1998), and Deniau et al. (2006).

The aim of our study is to illustrate and optimize existing methodology for the codominant scoring of AFLP markers using data from an association panel, without a priori knowledge of allele frequencies. The association panel consists of 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 et al. (2008b). Commercially available software, such as Quantar Pro (Keygene products BV 2004) from the private company Keygene NV, 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, 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 three 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.

Materials and methods

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 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 the intensities row-wise (per marker) by the maximum intensity per row, 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, such as 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 (three 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 and 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, ..., y_n. The Gaussian mixture model (McLachlan and Peel 2000) for intensity y_i of variety i is:

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

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 three 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

$$\tau_{ik} = \frac{\pi_k f_k(y_i)}{\sum_{j=1}^{3} \pi_j f_j(y_i)}, \qquad (2)$$

which are conditional genotype probabilities given the marker phenotype (intensity). In total, eight 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 Jansen (1993) and Piepho and Koch (2000). 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 2 of the next section. Mixture models are a topic of ongoing statistical research, because problems exist with the identifiability of parameters, and parameters occurring at the boundary of the parameter space. Therefore, most classical asymptotic results cannot be directly applied. Here, we supply a short review of recent work on mixture models. Böhning et al. (2007) give in an editorial an outline of the current state of the art. Slightly older is the book by McLachlan and Peel (2000), containing a wealth of references. Particularly, interesting aspects of mixture modeling for our situation are: (1) hypothesis testing, (2) order selection, i.e., determination of the number of groups, (3) robustification. Recent work on hypothesis testing for the special case of testing homogeneity (i.e., discriminating a one-component from a two-component mixture) is Chen and Li (2009), Li et al. (2009), and Garel (2007). The case of testing homogeneity is not of great interest in our situation, though. Other work focuses on testing homoscedastic versus heteroscedastic normal mixtures (e.g., Lo 2008), but conclusions are meager. The topic of order selection has kept statisticians busy for long. A worthwhile reference on hypothesis testing for the number of components is Feng and McCulloch (1994), but they describe the case of unequal variances, which we avoid (see following section). A very recent study on order selection is Chen and Khalili (2008) using a penalized likelihood approach. Comparing with other criteria in a simulation study, they conclude that their approach performs generally but not always better. Normality-based methods for estimation have the problem of sensitivity to outliers. Different authors studied the problem. Recent studies are McLachlan et al. (2006), using mixtures of t distributions, and Cuesta-Albertos et al. (2008), using a mix of initial robust clustering for subsamples and maximum likelihood. From this overview we learn that the final word on these topics has not been said.

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 2.
   (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 two 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 μs, 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 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 2. This could be done in different ways:
   (a) Transformation of band intensity
   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. Piepho and Koch (2000) study (Box-Cox) power transformations of the band intensity, optimizing the power by maximum likelihood to achieve normality.
   (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 and Nelder 1989). We will not pursue this topic further.

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 three components, allowing for 0, 1, or 2 copies of a DNA fragment. We may, however, face situations with only two 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 three 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 three 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, pp. 18–19) mentions that, for a two-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 four standard deviations or more the information is almost complete. To check the separation of groups, we propose to calculate for each AFLP marker $\text{sep}_1 = (\hat{\mu}_2 - \hat{\mu}_1)/\hat{\sigma}$ and $\text{sep}_2 = (\hat{\mu}_3 - \hat{\mu}_2)/\hat{\sigma}$ in the three-component normal mixture model with constant standard deviation σ. We call the separation "poor" if $\text{sep}_1 \leq 2$ or $\text{sep}_2 \leq 2$, "moderate" if not "poor", but $2 < \text{sep}_1 \leq 4$ or $2 < \text{sep}_2 \leq 4$, and "good" if $\text{sep}_1 > 4$ and $\text{sep}_2 > 4$. 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. We use two simple approaches: (1) identify outlying observations by simple visual inspection of the histogram (see item 2), 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 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 "Codominant scoring of AFLP band intensities by normal mixture models"), 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 to $\approx 10^6$), 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, as earlier described. We take 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_1^2 distribution of the test statistic $\text{LR} = 2(\text{LL(FM)} - \text{LL(RM)})$, 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 LRTs in mixture models, we underpin this approach by a small simulation study. We simulate band intensities for 100 individuals, by sampling from a three-component normal mixture with means $\mu =$ 0.3, 0.5, 0.7, a range of standard deviations σ = 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 "Data: association panel of tomato hybrids"). For the simulation, we first sample the genotypes of 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.

If $p \leq 0.2$, sets of genotypes may be sampled with $k_1 = 0$ (roughly 38% if $p = 0.1$, and 1.7% if $p = 0.2$), which we discard, as we would do for real data. In these cases we are sampling from a truncated multinomial distribution. Given the genotypes, we sample k_i intensities from $N(\mu_i, \sigma^2)$. From the fitted full and reduced models LR is calculated, and compared to the 95% critical value 3.84 of the χ_1^2 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 $\mathrm{BIC} = -2\mathrm{LL} + d \times \ln(n)$ 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 BICs. If the compared models have equal numbers of parameters, the comparison by BIC is equivalent to the comparison by LL.

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 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 Fig. 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 three 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). The resulting posterior genotype probabilities can be used in association mapping, analogously to the use of conditional QTL genotype probabilities given flanking marker information in case of QTL linkage mapping for biparental crosses.

Predictors for additive and dominance effects

Given the three 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 association mapping, relating the codominant scores to phenotypic information by mixed models. A paper on genome-wide association mapping using these scores is in preparation.

Data: association panel of tomato hybrids

Within the Centre for BioSystems Genomics, a Dutch plant genomic initiative (van Berloo et al. 2008a), 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 (van Berloo et al. 2008b). This set consists of 20 beef,

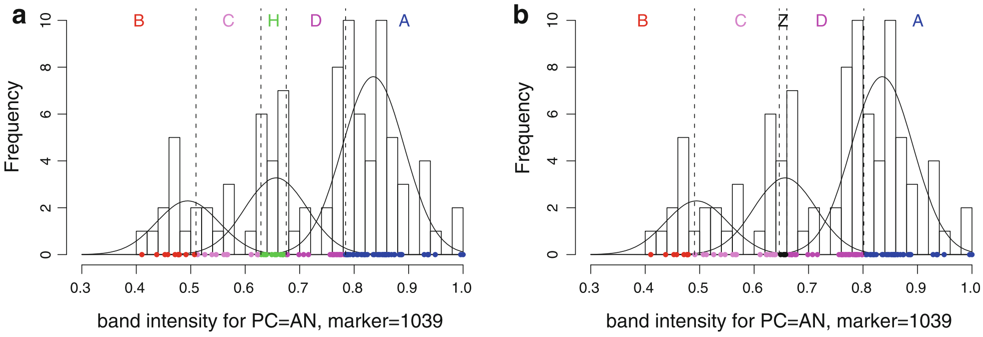

Fig. 1 Histograms of band intensities of marker 1,039 with superimposed normal densities. Subplots **a** and **b** show color-coded hard classifications based on probability thresholds 0.95, and 0.98, respectively. In the last case, some observations are classified as unknown (Z)

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 1,175 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 to $\approx 10^6$, and corrected intensities with values in the range $0-1$. We refer to the dataset of band intensities of 1,175 AFLP markers on 94 tomato hybrids as the “tomato data”.

Studying the scoring features in the complete tomato dataset

We study how the features mentioned in “Features for enhanced and stabilized unmixing, data quality and model selection” help in the codominant scoring of all 1,175 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 Jansen et al. (2001), in two ways:
   (a) Comparing the fits of homoscedastic and heteroscedastic three-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 three-component mixture model, is called best with respect to order selection.
3. Diagnostics for data quality of the 1,175 AFLP markers:
   (a) number of components: compare g-component homoscedastic mixture models (with $g = 2, 3, 4, 5$ components, and $d = 4, 6, 8, 10$ parameters, respectively) by BIC;
   (b) separation: count how often separation is poor, moderate or good in the best-fitting g-component model;
   (c) outliers: count how often singletons exist in the first or last component in the best-fitting g-component 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 “Features for enhanced and stabilized unmixing, data quality and model selection”. We use a selection of 300 mapped markers, following the paper by van Berloo et al. (2008b). Out of the 797 unmapped markers, we select 349 with best fitting three-component mixture model.

Results

Software

We developed software routines in R (Ihaka and 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 Materials and methods, 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". All plots and mixture model output in this paper are results from applications of the R routines.

Examples

Examples with well fitting mixture models

In Fig. 2, we show some examples of codominantly scored AFLP markers with well fitting three-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 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 "Features for enhanced and stabilized unmixing, data quality and model selection". 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 vs. $d = 6$ 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 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_js (LL = 21.9). This is an example of a pathological situation, because the EM-algorithm converges to an inferior solution for the full

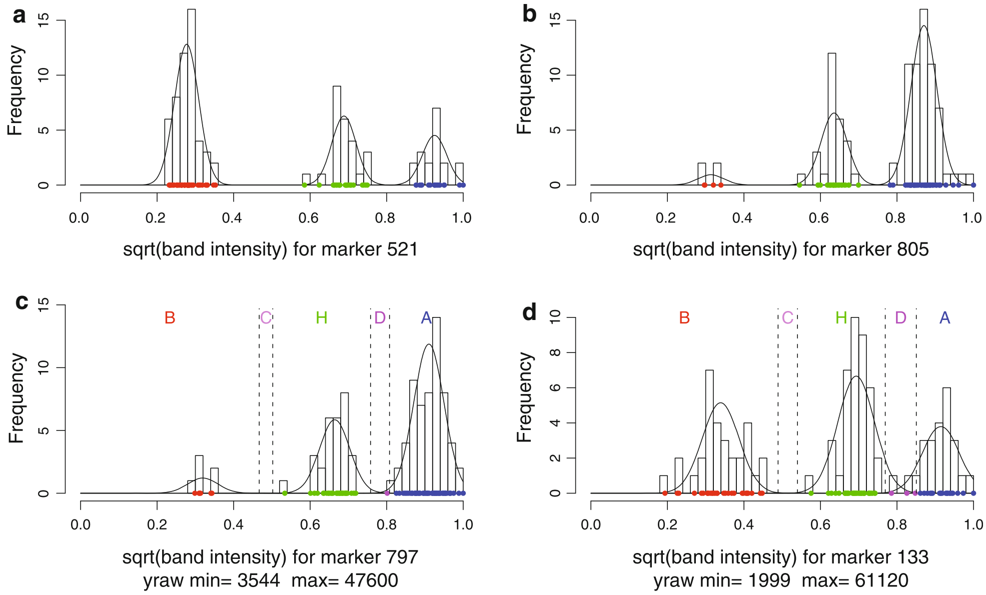

Fig. 2 Four examples of AFLP markers from the tomato data with histograms of band intensities, and well fitting normal mixture densities

model (free μ_js) 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 3a1–3a4 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_js. 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_js 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 four-component mixture, compared to three components, according to BIC.

   (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

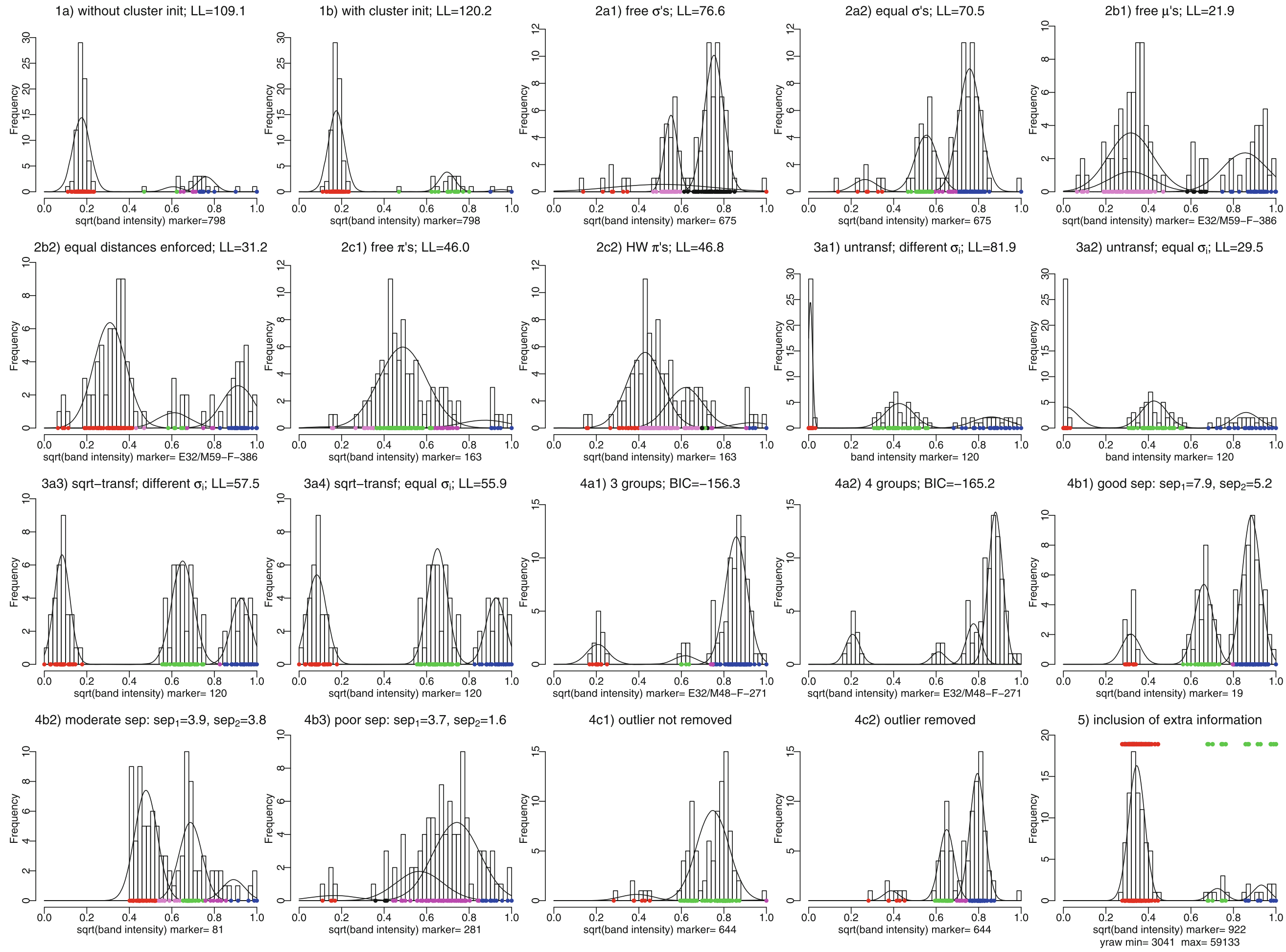

Fig. 3 Examples of features helping unmixing of marker intensities for the tomato data. Subplots **1a–b** deal with starting values of parameters; **2a1–a2** restriction on σ: hetero- versus homoscedasticity; **2b1–b2** restriction on μ: equidistant component means; **2c1–c2** HWE restriction on π; **3a1–a4**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

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).

Results for the complete tomato dataset

Parameter initialization

Table 1 shows the comparisons of the two types of parameter initialization of the EM-algorithm (by guesstimates and hierarchical clustering) for two-, three-, four-, and five-component homoscedastic mixture models for all 1,175 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.

Transformation of band intensity

Table 2 shows the comparison of homoscedastic and heteroscedastic three-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.

Table 1 Comparison of parameter initialization by log-likelihood of fitted models: guesstimates versus hierarchical clustering

	Number of groups			
	2	3	4	5
No difference	1,118	870	651	405
Guesstimate best	30	73	92	142
Cluster best	27	232	432	628
Total	1,175	1,175	1,175	1,175

Table 2 Comparison of homoscedastic and heteroscedastic three-component mixture models by BIC for a range of power transformations of band intensities. Percentages of markers with the homoscedastic model selected as best

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 3 shows the results of the comparisons of two-, three-, four-, and five-component homoscedastic mixture models for a range of power transformations. We find some very distinctive patterns. If the square root transformation is used, the three-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 2 and 3 that the square root transformation is best, both for variance stabilization and for order selection.

Diagnostics of data quality

Table 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 three components fits best for 561 markers ($\approx$50%). For 158 markers, a model with two components fits best. Models with more than three components are chosen for 456 markers. Results on the separation of group means in the best-fitting g-component model are shown in the middle part of Table 4. Notice that the majority of the markers (69%) have well separated group means, 31% is moderately separated, and only one marker is poorly separated. The percentages well separated markers monotonically decrease with the order g of the model: 89, 80, 53, and 34%, respectively. We conclude that the separation of group means shows a relationship with the choice of best fitting model.

The bottom part of Table 4 shows counts of markers with singletons in the last and first component of the best fitting g-component mixture model ($g = 2, 3, 4, 5$).

Table 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 1,175 are shown with a g-component normal mixture model (g = 2, 3, 4, 5) selected as best

g	Power transformation									
	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	1,175	1,175	1,175	1,175	1,175	1,175	1,175	1,175	1,175	1,175

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 the best fitting model: for markers with a best fitting three-component model, only 42 (7.5%) have a singleton in the last component, whereas markers with best fitting two-, four-, and five-component mixture models have singletons in 25, 26, and 36% of the cases, respectively.

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 three-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 two-, four- or five-component models suggest that the intensity is erroneous outlying (whatever the reason may be), and need further examination.

Table 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

	Number of components				Total
	2	3	4	5	
Selected as best	158	561	332	124	1,175
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

Testing for mixing probabilities according to Hardy–Weinberg equilibrium

Table 5 shows the results of the simulation study to underpin the LRT for HWE, as described in "Features for enhanced and stabilized unmixing, data quality and model selection". 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.

Figure 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 $P(\chi_1^2 \geq 0.8) = 0.37$. Hence, the null hypothesis of HWE is not rejected.

Table 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 three-component normal mixture model with means 0.3, 0.5, 0.7, using 10,000 replicates

σ	Allele frequency p				
	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

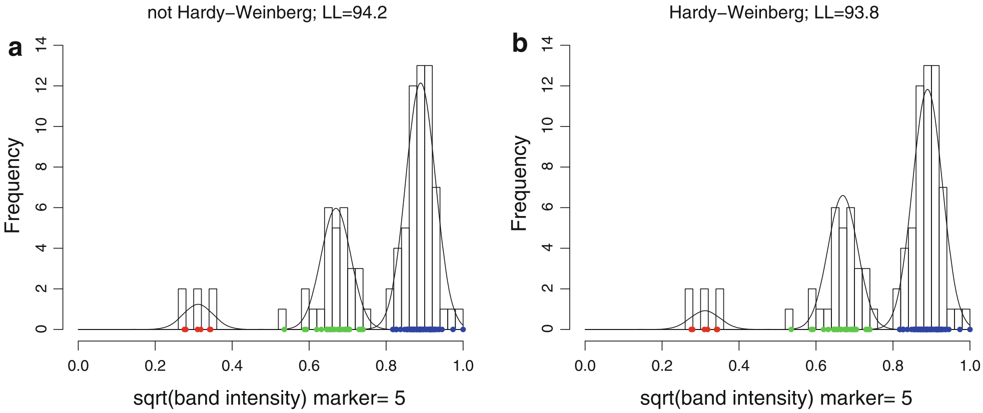

Fig. 4 Histogram and fitted normal mixtures with unrestricted π_j (subplot **a**) and restricted π_j according to HWE (subplot **b**)

The results for all selected markers are shown in Table 6 (cf. Table 2 in van Berloo et al. 2008b). If the LRT gives a P value of >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.

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", "B", "H", 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, depending on the purpose of the subsequent analysis. For example, in standard QTL mapping a marker label "C" or "D" may be changed into in informative label "A", "H", "B", using information from flanking markers. This is not possible in association mapping, where only information on the marker itself is used. We 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.

The EM-algorithm necessarily converges to a local maximum of the likelihood. Recently, papers appeared describing attempts for global optimization of the likelihood, using

Table 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

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

methods from Operations Research (Heath et al. 2009; Jank 2006a, b). Heath et al. (2009) mention that repeat application of EM (as we propose here) may achieve similar results. A further study into the global optimization of the likelihood in mixture models is advisable.

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 three components, valid for diploid organisms, is selected by BIC for about half of the 1,175 polymorphic bands (if choosing from models with 2, 3, 4, or 5 components). A model with more than three components is selected for about 38% of the markers. Models with more than three 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 four- or five-component model cannot be ruled out. A model with two 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-component 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 two-, four-, or five-component models have in 25–35% of the cases a single observation assigned to the component with highest mean, whereas for markers with best fitting three-component 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 two-, four- or five-component mixture models selected as best, contain more often an erroneous outlying observation than markers with three-component 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 three-component mixture model according to BIC, good separation of components, lack of singletons, robustness against parameter initialization, and robustness against slight data transformation. We have seen that many markers do not show the preferred number of three clusters, or have other characteristics that make them less optimal. An interesting question is what should be done with these markers. We do not recommend to discard these markers blindly, but instead use map information to decide on their use. If it concerns a mapped marker with many other neighbouring markers, it could easily be discarded. If the map is rather sparse, it may be worthwhile to check what is causing the problem.

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 6–18% (chromosomes 4 and 5) to 95–100% (chromosomes 3, 6, and 9). These differences may be caused by population substructure in the set of tomato cultivars. We found that chromosomes 4 and 5 contain markers related to the cherry/non-cherry subgroups.

Codominant scoring can also be exploited in AFLP mapping studies. AFLP maps are almost always based on dominantly scored markers. Piepho (2001) describes how band intensities can be used to infer the recombination frequency, and next to order the markers on a map. The information of band intensities is used by Pérez-Enciso and Roussot (2002) in a general pedigree to estimate identity by descent probabilities, to be used in subsequent QTL mapping strategies. 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 and Clarke 2007). An example of a study of this type of codominance is Wong et al. (2007).

Liu (2007) urges caution in the use of codominant scoring because of the non-linear 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.

Acknowledgments 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).

Appendix

Software description

We wrote software routines for the codominant scoring of AFLP profiles in R (Ihaka and 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 to 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 top of 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.

References

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

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

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

Böhning D, Seidel W, Alf M, Garel B, Patilea V, Walther G (2007) Advances in mixture models. Comput Stat Data Anal 51:5205–5210

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. Theor Appl Genet 99:425–431

Chen J, Khalili A (2008) Order selection in finite mixture models with a nonsmooth penalty. J Am Stat Assoc 103:1674–1683

Chen J, Li P (2009) Hypothesis test for normal mixture models: the EM approach. Ann Stat 37:2523–2542

Cuesta-Albertos J, Matrán C, Mayo-Iscar A (2008) Robust estimation in the normal mixture model based on robust clustering. J Roy Stat Soc B Met 70:779–802

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via EM algorithm. J Roy Stat Soc B Met 39:1–38

Deniau AX, Pieper B, Ten Bookum WM, Lindhout P, Aarts MGM, Schat H (2006) QTL analysis of cadmium and zinc accumulation in the heavy metal hyperaccumulator *Thlaspi caerulescens*. Theor Appl Genet 113:907–920

van Eck HJ, Rouppe van der Voort J, Draaistra J, van Zandvoort P, van Enckevort E, Segers B, Peleman J, Jacobsen E, Helder J, Bakker J (1995) The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring. Mol Breed 1:397–410

Feng ZD, McCulloch CE (1994) On the likelihood ratio test statistic for the number of components in a normal mixture with unequal variances. Biometrics 50:1158–1162

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97:611–631

Garel B (2007) Recent asymptotic results in testing for mixtures. Comput Stat Data Anal 51:5295–5304

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

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

Heath JW, Fu MC, Jank W (2009) New global optimization algorithms for model-based clustering. Comput Stat Data Anal 53:3999–4017

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314

Jank W (2006a) Ascent EM for fast and global solutions to finite mixtures: An application to curve-clustering of online auctions. Comput Stat Data Anal 51:747–761

Jank W (2006b) The EM algorithm, its randomized implementation and global optimization: some challenges and opportunities for operations research. In: Alt FB, Fu MC, Golden BL (eds) Perspectives in operations research; papers in honor of Saul Gass' 80th birthday, Chap. 21, part III. Springer, US, pp 367–392

Jansen RC (1993) Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. Biometrics 49:227–231

Jansen RC (1994) Maximum likelihood in a finite mixture model by exploiting the GLM facilities of Genstat. Genstat Newsl 30:25–27

Jansen RC, Geerlings H, van Oeveren AJ, van Schaik RR (2001) A comment on codominant scoring of AFLP markers. Genetics 158:925–926

Keygene Products BV (2004) AFLP-Quantar(r)Pro 1.0. Part I—user guide. Keygene Products NV, Wageningen

Li P, Chen J, Marriott P (2009) Non-finite Fisher information and homogeneity: an EM approach. Biometrika 96:411–426

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

Liu Z (2007) Amplified fragment length polymorphism (AFLP). In: Liu Z (ed) Aquaculture genome technologies, Chap 4. Blackwell, Ames, pp 29–42

Lo Y (2008) A likelihood ratio test of a homoscedastic normal mixture against a heteroscedastic normal mixture. Stat Comput 18:233–240

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London

McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York

McLachlan GJ, Ng S, Bean R (2006) Robust cluster analysis via mixture models. Austrian J Stat 35:157–174

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

Pérez-Enciso M, Roussot O (2002) A method for computing identity by descent probabilities and quantitative trait loci mapping with dominant (AFLP) markers. Genet Res 79:247–258

Piepho HP (2001) Exploiting quantitative information in the analysis of dominant markers. Theor Appl Genet 103:462–468

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

Reamon-Büttner SM, Schondelmaier J, Jung C (1998) AFLP markers tightly linked to the sex locus in *Asparagus officinalis* L. Mol Breed 4:91–98

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

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

Wong A, Forbes MR, Smith ML (2007) Characterization of AFLP markers in damselflies: prevalence of codominant markers and implications for population genetic applications. Genome 44:677–684