Codominant Analysis of Banding Data From a Dominant Marker System by Normal Mixtures

Hans-Peter Piepho* and Georg Koch†,1

**Institut für Nutzpflanzenkunde, Universität-Gesamthochschule Kassel, 37213 Witzenhausen, Germany and †Institut für Pflanzenbau und -züchtung, Universität Kiel, 24118 Kiel, Germany*

Manuscript received November 30, 1999
Accepted for publication April 3, 2000

ABSTRACT

Amplified fragment length polymorphisms (AFLPs) currently are among the most widely used marker systems. In many studies, AFLPs are analyzed on the basis of the presence or absence of a band on an electrophoretic gel. As a result, dominant homozygous individuals are not distinguished from heterozygous individuals, resulting in a considerable loss of information. This article shows how codominant information can be obtained if the amount of PCR products is quantified. Due to measurement variation, genotyping on the basis of such information is not error-free. We propose use of normal mixture distributions to determine the most likely genotype, given the data. The method is exemplified using AFLP data from sugar beet.

MOLECULAR markers are increasingly being used in plant breeding research. Codominant markers are preferable to dominant markers due to the larger information content. A codominant molecular marker allows unequivocal distinction of homozygous and heterozygous genotypes on an electrophoretic gel. By contrast, for dominant markers, dominant homozygous and heterozygous individuals cannot be distinguished on the basis of the presence or absence of bands on a gel. Nevertheless, dominant markers continue to be very popular, mainly because of economical reasons. In this article, we consider the analysis of amplified fragment length polymorphisms (AFLPs), a marker system that has been used in many studies. In the past, numerous articles have analyzed AFLPs on the basis of band presence/absence, extracting only dominant information. The present article suggests a codominant analysis, which allows far more information to be exploited. We expect this method to be of wide interest to plant breeders and geneticists, since it makes more efficient use of dominant marker data.

In principle, the genotype of a dominant marker can be inferred from the optical density (OD) of the band on the gel or from the fluorescence in a gel-free marker assay. For simplicity, both types of data are henceforth referred to as OD values. For homozygous individuals, the OD value is expected to be larger than for heterozygous individuals, since for the latter the amount of PCR products should be only half that of homozygous individuals. Thus, if the OD can be measured quantitatively, there is a basis for identifying the genotype, *i.e.*, to extract the full codominant information.

Due to various sources of random variation (marker assay, measurement errors, gel differences, differences in duration of development, and concentration of developer, etc.), however, the measurements vary among individuals even if they have the same genotype, and a distinction among genotypes is not generally error free. The error variance may be so large that distributions of the three genotype classes overlap. Estimation of genotype frequencies in the population and of genotyping is therefore not straightforward and use of statistical methods may be helpful.

In this article, we specifically consider the following problem: A PCR-based dominant marker system (AFLP or other) is used to score a population of plants for their genotype at the marker locus. For each individual, the concentration of PCR products at a band position on an electrophoretic gel is measured quantitatively, either directly as pixel information or indirectly via an OD value. There are three genotypes, which may be denoted A_1A_1, A_1A_2, and A_2A_2, where A_1 is the null allele and A_2 is the dominant allele. The objectives are (1) to assign each individual to one of the genotype classes and (2) to estimate the frequencies of the three genotypes in the population. If genotype frequencies are known, only task (1) remains. For example, for a segregating F_2 population it may be known in advance that the frequencies are 0.25, 0.50, and 0.25 for A_1A_1, A_1A_2, and A_2A_2, respectively. This article discusses application of normal mixture models (Everitt and Hand 1981; McLachlan and Basford 1988; Hastie and Tibshirani 1996;

This article is dedicated to Prof. Dr. M. Hühn (Institut für Pflanzenbau- und züchtung, Universität Kiel, Germany) on the occasion of his 60th birthday.

Corresponding author: Hans-Peter Piepho, Institut für Nutzpflanzenkunde, Universität-Gesamthochschule Kassel, Steinstrasse 19, 37213 Witzenhausen, Germany. E-mail: piepho@wiz.uni-kassel.de

[1] *Present address:* A. Dieckmann-Heimburg, Postfach 1165, 31684 Nienstädt, Germany.

Lynch and Walsh 1997) for these purposes. The methods are exemplified using real data on AFLP markers in a segregating F_2 population of sugar beet.

THEORY

It can be assumed that conditionally on the genotype class the observed values for individual plants follow some continuous distribution. The joint distribution of individuals from all three genotype classes is then a mixture of three distributions, with mixing proportions equal to the genotype frequencies. If genotype frequencies are unknown, they can be estimated by estimates of the mixing proportions. Allele frequencies can be estimated from the genotype frequencies in the usual way. Using the estimated distribution of the data, individuals may be assigned to one of the three genotype classes based on a *posterior* probability of genotype class membership.

One possible assumption is that OD values for individuals from a given genotype class are distributed normally. This assumption can and should be checked if possible. The joint distribution is then a mixture of three normal distributions. Let g be a random variable denoting the genotype class with $g = 1$ when the genotype is A_1A_1, $g = 2$ when the genotype is A_1A_2, and $g = 3$ when the genotype is A_2A_2 (for simplicity we make no distinction in notation between a random variable and its realization). Group membership of a randomly drawn individual from the population follows a multinomial distribution with parameters $\pi_i = P(g = i)$ and index 1. The G-component normal mixture probability density function (p.d.f.) of y_j, the OD value of the jth plant ($j = 1, \ldots, n$) can be written

$$f(y_j|\theta) = \sum_{i=1}^{G} P(g = i)\phi(y_j|\mu_i, \sigma_i), \qquad (1)$$

where $\theta = (\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, \pi_1, \pi_2, \pi_3)'$, $\pi_i = P(g = i)$ are the nonnegative mixing proportions, subject to the constraint $\Sigma_{i=1}^{G}\pi_i = 1$, and $\phi(y_j|\mu_i, \sigma_i)$ is a normal distribution with mean μ_i and standard deviation σ_i. Here, we have $G = 3$.

According to Bayes' theorem (Cox and Hinkley 1974), the posterior probability of genotype class membership of an individual, given its phenotypic value y_j, is

$$P(g = i|y_j;\theta) = \frac{P(g = i)\phi(y_j|\mu_i,\sigma_i)}{f(y_j|\theta)}, \qquad (2)$$

where $\pi_i = P(g = i)$ plays the role of a *prior* probability of genotype class membership. The posterior probabilities may be evaluated by replacing parameters with their estimates. On the basis of their estimated posterior probabilities, individuals may be assigned to one of the three genotype classes (genotyping). In the appendix, we describe how to compute the correct allocation rate (CAR), *i.e.*, the probability that a randomly selected individual is correctly classified.

There are two possible simplifications of the model:

1. The mixing proportions $\pi_i = P(g = i)$ are known *a priori*. For example, in a segregating F_2 population the proportions may be known to be $\pi_1 = 0.25$, $\pi_2 = 0.50$, $\pi_3 = 0.25$. This is the case for the example considered in the next section.
2. The variances of the three components are equal.

Hence, we consider four different models:

a. $\theta = (\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3, \pi_1, \pi_2, \pi_3)'$ is completely free (apart from the usual boundary constraints $\sigma_i > 0$, $\pi_i > 0$, and $\pi_1 + \pi_2 + \pi_3 = 1$). There are thus eight free parameters (only two of the three mixing proportions are free to vary).
b. $\sigma_1 = \sigma_2 = \sigma_3 = \sigma$. There are six free parameters.
c. $\pi_1 = 0.25$, $\pi_2 = 0.50$, $\pi_3 = 0.25$. There are six free parameters.
d. $\sigma_1 = \sigma_2 = \sigma_3 = \sigma$ and $\pi_1 = 0.25$, $\pi_2 = 0.50$, $\pi_3 = 0.25$. There are four free parameters.

To fit these models, we use the expectation-maximization (EM) algorithm. A brief description is given in the appendix. The models (a)–(d) may be compared by likelihood-ratio tests. Let θ_F be the parameter vector for a full model and θ_R the parameter vector of a reduced model relative to θ_F. We are interested in testing the null hypothesis, that $\theta_F = \theta_R$. Under H_0 $\theta_F = \theta_R$, the statistic

$$T = -2[\log L(\hat{\theta}_R) - \log L(\hat{\theta}_F)],$$

where $L(\cdot)$ is the likelihood function, $\hat{\theta}_F$ and $\hat{\theta}_R$ are the maximum-likelihood estimates of θ_F and θ_R, respectively, is asymptotically (*i.e.*, in large samples) distributed as χ^2 with q d.f., where q is the difference in the number of free parameters between θ_F and θ_R. To test for significant variance homogeneity, we compare (a) (full model) *vs.* (b) ($q = 2$). To test for significant departure from the segregation rate $\pi_1 = 0.25$, $\pi_2 = 0.50$, $\pi_3 = 0.25$, we compare model (a) (full model) *vs.* model (c) ($q = 2$). If both tests are significant, we select model (a). If only the first is significant, we choose model (c). If only the second is significant, we choose model (b). If neither test is significant, we consider model (d) as the most appropriate model. Since two independent tests are conducted per band position, we control the family-wise error rate at α by performing each test at a significance level of $\alpha/2$ (this is a Bonferroni procedure; see Hochberg and Tamhane 1987).

So far, we have assumed normality of the data. It is not certain that this assumption will always be met in practice. In principle, mixture distributions with components other than normal distributions can be contemplated (Redner and Walker 1984), but then the problem arises as to the choice of distribution. An alternative is to seek a normalizing transformation of the data.

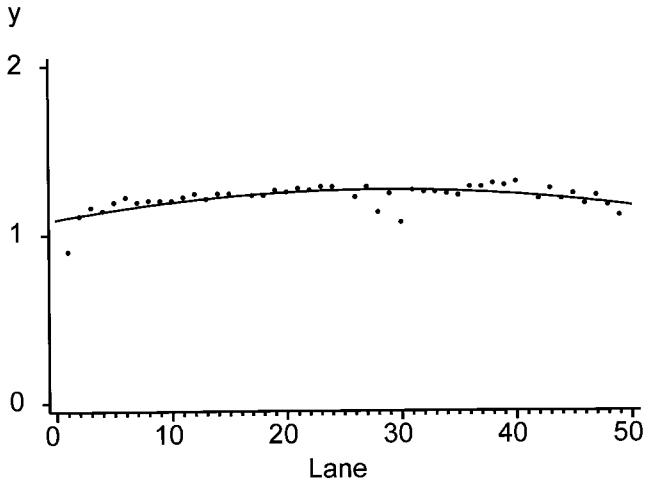

Figure 1.—Plot of OD values (y) *vs.* lane number for a monomorphic band. Fitted line is a polynomial of the form $\mu + \beta_1 x_j + \beta_2 x_j^2$.

Gutierrez *et al.* (1995) discuss application of the Box and Cox (1964) power transformation to normal mixtures. This family of transformations is very flexible and includes the logarithmic transformation as a special case. It is given by

$$h(y;\lambda) = y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0. \end{cases}$$

Note that taking $\lambda = 1$ is equivalent to not transforming the data. The simplest way to obtain a maximum-likelihood estimate of λ is by a grid search (for details see Gutierrez *et al.* 1995). Once an appropriate transformation has been found, a normal mixture may be fitted to the transformed data.

EXAMPLES

As an example we used OD values at 13 band positions of an AFLP marker for 46 individual F_2 plants (sugar beets). OD values were preprocessed to subtract systematic trends discernible from four monomorphic band positions. Figure 1 shows a plot of OD values for one of these monomorphic band positions *vs.* lane number on the gel. There is a clear nonlinear trend. To remove trends, a quadratic polynomial regression model was fitted to each monomorphic band position. Using the data of the four band positions, we performed a joint analysis of covariance based on the model $y_{ij} = \mu_i + \beta_1 x_j + \beta_2 x_j^2 + \delta_{1i} x_j + \delta_{2i} x_j^2 + a_j + e_{ij}$, where y_{ij} = OD value for the jth lane and the ith band position, x_j = lane position number (centered at mean lane number of 25), and a_j = random lane effect with zero mean and variance σ_a^2. To check for spatial correlation among adjacent lanes, we fitted two models for a_j: (i) independent effects a_j and (ii) effects correlated according to a Gaussian model, $\text{cov}(a_s, a_t) = \sigma_a^2[\exp(-d_{st}^2/\rho^2)]$, where d_{st} is the absolute difference in lane position numbers among lanes s and t. Both models were fitted by restricted maximum likelihood (REML) using the MIXED procedure of the SAS System (SAS Institute, Inc., 1997 SAS/STAT software, changes and enhancements through release 6.12). We computed the Schwarz Bayesian criterion (SBC) for both models as (i) SBC = 162.82 and (ii) SBC = 161.54. The larger SBC for model (i) indicates that there is little evidence of residual spatial correlation (Wolfinger 1996).

An analysis of variance based on model (i) for a_j (Table 1) showed that common slopes β_1 and β_2 can be assumed for all four band positions. Thus, it was concluded that the regression can be used for correcting the OD values of all other band positions. The correction term subtracted from all OD values was $c = \beta_1 x + \beta_2 x^2$. The regression terms were estimated as $\beta_1 = 0.000912310$ and $\beta_2 = -0.000228513$ by fitting the model $y_{ij} = \mu_i + \beta_1 x_j + \beta_2 x_j^2 + a_j + e_{ij}$ jointly to the four monomorphic bands, where μ_i is an intercept term corresponding to the ith band and a_j is a random effect of the jth lane. Since the REML analysis of models (i) and (ii) revealed no evidence of spatial correlation, corrected OD values were regarded as stochastically independent in the subsequent analyses. We also considered the possibility of correcting OD values separately for each individual using monomorphic band positions. This option was discarded, however, due to low correlation among different monomorphic band positions across individuals.

Figure 2 shows the detrended OD values at the 13 polymorphic band positions for 46 individual plants. The expected segregation at a locus is 1:2:1. For some band positions, the null bands (genotype A_1A_1) appear to be clearly discernible from the other alleles (*e.g.*, band positions 2, 5, 9, and 13), while for others the distinction is not so clear. In all cases, the separation between the heterozygous and the homozygous dominant genotypes is not clear cut. Our purpose is to initially fit models (a)–(d) and subsequently select an appropriate model for genotyping. The log-likelihoods of models (a)–(d) as fitted to data from each of these 13 band positions are shown in Table 2. For band positions 2, 3, 7, and 12, there is a significant departure from the expected diallelic segregation rate ($\pi_1 = 0.25$, $\pi_2 = 0.50$, $\pi_3 = 0.25$). Also, band positions 2, 3, 4, 6, 11, and 12 show significant heterogeneity of variance. Under model (a), estimated variances often differ by several decimal places (see Table 3). Noteworthy examples are bands 6, 9, 11, and 12, where one group has a strikingly small mixing proportion, and this is associated with a conspicuously small variance. For band positions 1, 5, 8, 9, 10, and 13, the simultaneous assumptions of homogeneous variances and of diallelic segregation are tenable according to the likelihood-ratio tests. Note that the sample size ($n = 46$ plants) is rather small, so the results

TABLE 1

Analysis of variance for regression of (uncorrected) OD value on lane number

Source	d.f.	MS[a]	*F*	*P* value
Band (μ_i)	3	3.77	1208	<0.0001
Linear ($\beta_1 x_j$)	1	0.0348[b]	2.79	0.1090
Quadratic ($\beta_2 x_j^2$)	1	0.315[b]	25.26	<0.0001
Band*linear ($\delta_{1i} x_j$)	3	0.00810	2.60	0.0553
Band*quadratic ($\delta_{2i} x_j^2$)	3	0.00133	0.43	0.7349
Lane (a_j)	43	0.0125	3.99	<0.0001
Error (e_{ij})	172	0.00312		

Band, band position; x_j, lane position number − 25.

[a] Mean squares based on sequential reduction in residual sums of squares due to fitting the term in question (type I). Fitting sequence is in the order of appearance of terms in the table.

[b] Tested against the mean square for lane.

of likelihood-ratio tests, which are valid only asymptotically, should just be taken as a rough guide.

As an example for genotyping we use band position 1. The data are shown in Table 4. Model (d) was used to compute posterior probabilities, since this fitted best according to likelihood-ratio tests (Table 2). The parameter estimates were $\hat{\mu}_1 = 0.371$, $\hat{\mu}_2 = 0.793$, $\hat{\mu}_3 = 1.026$, and $\hat{\sigma}^2 = 0.0107$. The resulting grouping is indicated by underscoring in Table 4. Figures 3 and 4 provide a graphical representation of the fitted normal mixture model. While the A_1A_1 genotype is relatively clearly removed from the rest of the data, the distributions for genotypes A_1A_2 and A_2A_2 show considerable overlap. The figures give an impression of how difficult it would be to distinguish genotypes A_1A_2 and A_2A_2 by eye. Also note that the abscissa value of the point of intersection among two component normal curves in Figure 3 is the class limit for plants/OD values classified as belonging to either the one or the other component. There were 16 plants genotyped as A_1A_1, 20 as A_1A_2, and 10 as A_2A_2. This compares favorably well with the expected segregation of 1:2:1. The example shows that the statistical method is a useful aid in distinguishing the genotypes. The correct allocation rate (see appendix) based on model (d) was estimated as CAR = 0.896; *i.e.*, ~90% of the individuals are expected to be classified correctly.

To check tenability of the normality assumption, the Box-Cox transformation was applied to model (d). The maximum-likelihood estimate of λ was 0.32. This is somewhat removed from λ = 1, which corresponds to the untransformed data. Note, however, that the associated value of the maximized log-likelihood of 0.876 is only marginally different from the value obtained for the untransformed data (0.143), so there is little evidence of nonnormality. The associated likelihood-ratio test for H_0: λ = 1 is not significant at α = 0.05. Thus,

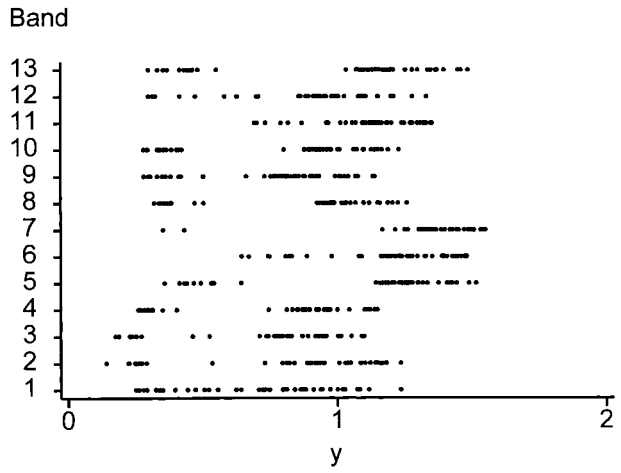

Figure 2.—Plot of OD values (*y*) for 46 individual plants at 13 band positions (BD). Some dots represent more than one observation.

TABLE 2

Log-likelihoods for models a–d

Model:	a	b	c	d
No. of parameters:	8	6	6	4
1	3.447	1.107	1.175	0.143
2	15.127	9.487	10.626	8.316
3	21.756	10.106	10.747	9.935
4	34.464	27.504	33.050	26.387
5	29.381	27.274	27.988	25.815
6	20.376	12.805	16.951	12.166
7	51.306	49.599	36.220	34.987
8	32.947	32.573	32.568	31.976
9	19.573	16.574	16.786	16.509
10	27.772	27.710	27.414	27.327
11	30.748	26.872	29.539	24.859
12	14.555	1.337	4.336	1.017
13	17.114	16.905	16.987	16.755

Model identified as appropriate by likelihood-ratio test is underlined. χ^2(d.f. = 2; α = 0.025) = 7.38; the critical likelihood difference in the log-likelihood values for comparison of model a *vs.* b and a *vs.* c is 3.69 (in these pairs of models the difference in the number of parameters is 2).

we conclude that fitting a normal mixture to the untransformed data is tenable.

WHY EXTREME OD VALUES SHOULD BE AVOIDED

The OD value observed for a given band is a function of the PCR product concentration at the band. It may happen that due to the measurement process, the OD value is bounded upward. In this case, the relationship between concentration and OD value should be nonlinear, with a horizontal asymptote at OD_{max} as the concentration of PCR products increases (see examples in Figure 5). It is clear that the information content of OD values decreases as OD_{max} is approached. For most measurement devices, the upper limit is well known or is easily determined experimentally. Wherever possible, extreme OD values due to, *e.g.*, overloading of the gel or overexposure of the X-ray film should be avoided. In what follows, we consider an experiment to study the effect of extreme OD values. Implications for a need to transform the data are discussed.

We ran an experiment with material from two sugar beet plants used in previous linkage mapping projects. The two plants were the parents of a cross used to derive a mapping population (F_2) for another study. From previous analyses, the two plants were known to exhibit polymorphisms for a number of band positions. Thus, the plants were known to be either A_1A_1 or A_2A_2 at a number of loci. DNA was extracted from both plants. The extracted DNA was mixed in different proportions (0–100% A_2A_2). The mixing proportion can be assumed to be proportional to the concentration of PCR products found on the gel. For each mixing proportion we generated AFLPs in two replications. For each replication, two samples of PCR product were subjected to electrophoresis. Thus, there were a total of four analyses per mixing proportion.

A total of 14 band positions were analyzed. To study the asymptotic behavior of OD values as the amount of PCR product increases, an exponential function of the form

$$y = A - B\exp(Cz) \tag{3}$$

was fitted to the resulting plot of y (OD) *vs.* mixing proportions z ($z = 0$–1 for $A_2A_2 = 0$–100%) by nonlinear least squares. Note that under model (3), y approaches an asymptote from below at $y = A$ as z approaches infinity ($C < 0$), so it is well suited to model asymptotic behavior. Figure 5 shows plots for two band positions, together with the fitted curves. The fit of the exponential function to these particular data was remarkably good. With most of the 14 band positions, an asymptote was approached with increasing z. The estimated asymptote was usually between 1.6 and 2. Inspection of the four replicates indicated consistent differences among band positions in the asymptotic value (results not shown).

The dotted OD curve in Figure 5 rapidly approaches OD_{max}, so distinction among A_2A_2 plants ($z = 1$) and A_1A_2 plants ($z = 0.5$) is difficult; in other words, the information content of the measured OD values is low. The curvature of the solid OD curve is less pronounced, and the relationship between y and z is nearly linear, so the distinction of A_2A_2 plants ($z = 1$) and A_1A_2 plants ($z = 0.5$) is more clear cut. The example stresses the need to tune the system so that all OD values remain in the medium range of possible values.

If the relationship between y and z is known to be $y = f(z)$, the inverse function $z = f^{-1}(y)$ can be used to infer z from y. It can also be conjectured that this inverse transformation achieves approximate normality if some of the observed values of y are close to the boundary. In the case of the exponential function (3), this requires estimation of the parameters A, B, and C using independent data from several controls running on the same

TABLE 3

Estimates of normal mixture with unknown mixing proportions and heterogeneous variances (model a)

Band position	μ_1	μ_2	μ_3	σ_1^2	σ_2^2	σ_3^2	π_1	π_2	π_3
1	0.300	0.468	0.866	0.00159	0.00245	0.0256	0.198	0.125	0.677
2	0.242	0.946	1.162	0.00169	0.0201	0.000182	0.196	0.664	0.141
3	0.234	0.840	1.094	0.000836	0.0173	0.0000141	0.239	0.680	0.081
4	0.296	0.899	1.122	0.00125	0.00454	0.000213	0.348	0.479	0.173
5	0.479	1.242	1.437	0.00627	0.00231	0.00207	0.196	0.629	0.175
6	0.643	0.785	1.295	0.324E-07	0.00472	0.0150	0.043	0.177	0.780
7	1.231	1.338	1.477	0.00177	0.000438	0.00207	0.092	0.479	0.385
8	0.376	0.995	1.179	0.00254	0.00159	0.00248	0.283	0.524	0.193
9	0.348	0.501	0.885	0.00214	0.36E-06	0.0156	0.195	0.043	0.761
10	0.349	0.929	1.138	0.00193	0.00232	0.00203	0.283	0.437	0.281
11	0.763	1.098	1.171	0.00396	0.303E-08	0.0118	0.124	0.043	0.833
12	0.437	1.000	1.210	0.0202	0.0208	0.666E-08	0.221	0.714	0.065
13	0.414	1.149	1.382	0.00454	0.00313	0.00428	0.283	0.480	0.238

Sugar beet data.

TABLE 4

Genotyping for band position 1 (model d)

Plant no.	y_j	Posterior probabilities $P(g = 1\|y_j; \theta)$	$P(g = 2\|y_j; \theta)$	$P(g = 3\|y_j; \theta)$	Inferred genotype[a]
1	0.50352	0.91890	0.08109	0.00001	1
2	1.02187	0.00000	0.14752	0.85248	3
3	1.00067	0.00000	0.21554	0.78446	3
4	0.24993	1.00000	0.00000	0.00000	1
5	0.72965	0.00145	0.98881	0.00975	2
6	0.51983	0.85592	0.14407	0.00001	1
7	0.84046	0.00002	0.90048	0.09950	2
8	0.44155	0.99247	0.00753	0.00000	1
9	1.08310	0.00000	0.04353	0.95647	3
10	1.07510	0.00000	0.05140	0.94860	3
11	1.11756	0.00000	0.02101	0.97899	3
12	0.91048	0.00000	0.66272	0.33728	2
13	0.33385	0.99989	0.00011	0.00000	1
14	1.02769	0.00000	0.13226	0.86774	3
15	0.72197	0.00196	0.98978	0.00825	2
16	0.62192	0.09435	0.90480	0.00085	2
17	0.79758	0.00010	0.95834	0.04156	2
18	0.55370	0.60837	0.39155	0.00008	1
19	0.34027	0.99986	0.00014	0.00000	1
20	0.93731	0.00000	0.52256	0.47744	2
21	0.39479	0.99881	0.00119	0.00000	1
22	0.97274	0.00000	0.33568	0.66432	3
23	0.81114	0.00005	0.94486	0.05508	2
24	0.71932	0.00218	0.99003	0.00779	2
25	1.01909	0.00000	0.15530	0.84470	3
26	0.86932	0.00000	0.82824	0.17176	2
27	0.64001	0.04843	0.95025	0.00133	2
28	0.44115	0.99259	0.00741	0.00000	1
29	0.94275	0.00000	0.49287	0.50713	3
30	0.79481	0.00011	0.96067	0.03922	2
31	0.74733	0.00072	0.98501	0.01427	2
32	0.47030	0.97686	0.02314	0.00000	1
33	0.74373	0.00083	0.98596	0.01321	2
34	0.87761	0.00000	0.80095	0.19904	2
35	0.88196	0.00000	0.78542	0.21458	2
36	0.70676	0.00359	0.99048	0.00592	2
37	0.86202	0.00001	0.84973	0.15026	2
38	1.23773	0.00000	0.00156	0.99844	3
39	0.29053	0.99998	0.00002	0.00000	1
40	0.27762	0.99999	0.00001	0.00000	1
41	0.25516	1.00000	0.00000	0.00000	1
42	0.32316	0.99993	0.00007	0.00000	1
43	0.26162	0.99999	0.00001	0.00000	1
44	0.80053	0.00008	0.95571	0.04420	2
45	0.86990	0.00000	0.82643	0.17357	2
46	0.34973	0.99980	0.00020	0.00000	1

Most probable genotypes are underscored.
[a] 1, A_1A_1; 2, A_1A_2; 3, A_2A_2.

gel as the genotypes to be scored, which is not usually feasible for routine work. The inverse of function (3) is linear in $\log(A - y)$, where $A = \mathrm{OD}_{\max}$. This suggests that we may use the transformation $\log(\mathrm{OD}_{\max} - y)$, providing $\mathrm{OD}_{\max}$ is known. This transformation is only useful for further analysis by normal mixtures, however, if the transformed data are more nearly normal than are the untransformed data. Also, it is not clear whether $\mathrm{OD}_{\max}$ is constant over band positions. We considered Box-Cox transformations of $(A - y)$ to achieve normality [note that $\log(A - y)$ is a special case]. In the case at hand, we observed $\mathrm{OD}_{\max} \approx 2$. Box-Cox transformation

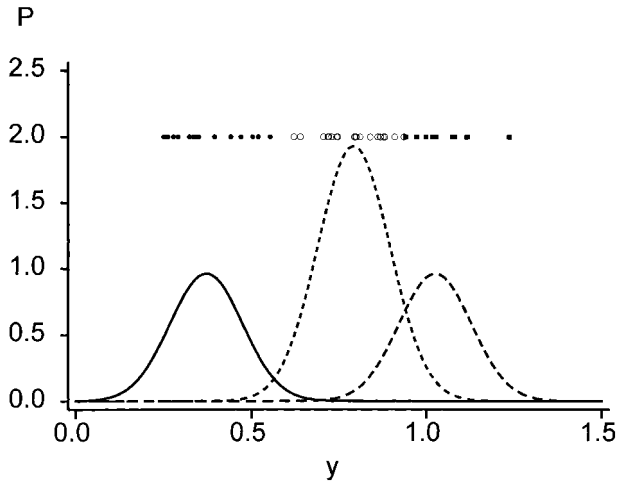

Figure 3.—Fitted normal mixture for band position 1 [model d; component normal distributions, weighted by $P(g = i)$]. Classification of OD values: A_1A_1, solid circles; A_1A_2, open circles; A_2A_2, solid squares. Component distribution is given by $\pi_i\phi(y_j|\mu_i, \sigma_i)$.

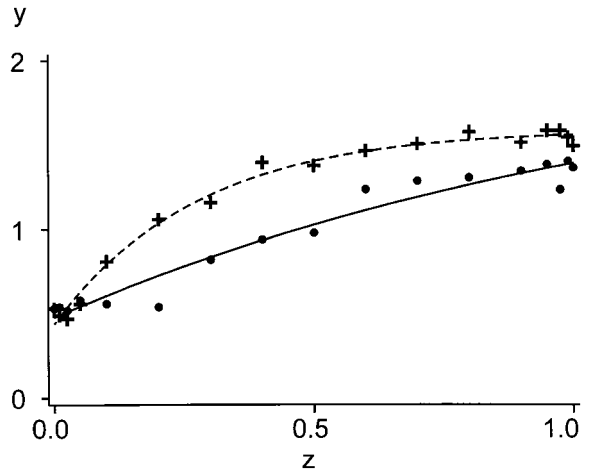

Figure 5.—Plot of OD value (y) *vs.* proportion of A_2A_2 (z) for two band positions. Dashed curve, $y = 1.60 - 1.16 \exp(-3.63z)$; solid curve, $y = 2.15 - 1.67 \exp(-0.80z)$.

of $(2 - y)$ for the data of band position 1 yielded an estimate of $\lambda = 1.45$ with a log-likelihood value of 0.281. Again, since the likelihood is only slightly different from that obtained for the untransformed data, this result is not indicative of nonnormality of $(2 - y)$ and hence of y. Thus, the untransformed data were used for analysis.

It can be conjectured that the expected concentration of PCR products (z) for the heterozygous plants is intermediate between the expected concentration of the two homozygous genotypes. Thus, we may contemplate the constraint $\mu_2 = (\mu_1 + \mu_3)/2$. This suggests that z values can be analyzed by a normal mixture with only two parameters for the means. Note that for our analyses of OD values (y), three independent means were estimated. Imposing the constraint $\mu_2 = (\mu_1 + \mu_3)/2$ is expected to avoid or largely decrease the likelihood of obtaining unrealistic fits, *e.g.*, a mixture in which two of the three component normal distributions have very similar means but widely different variances, a common problem when fitting normal mixtures. Again, however, this requires knowledge of the relationship $y = f(z)$, which is not usually available. Perusal of our data indicates that the simple transformation $z = \log(2 - y)$ does not generally lead to data commensurate with the constraint $\mu_2 = (\mu_1 + \mu_3)/2$. This issue will be pursued further in future work.

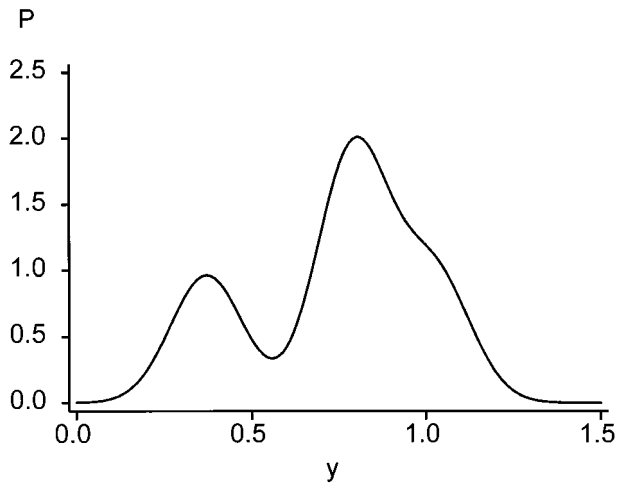

Figure 4.—Fitted normal mixture for band position 1 (model d).

DISCUSSION

The method proposed here is closely related to discriminant analysis (McLachlan and Basford 1988). The main difference lies in the fact that discriminant analysis uses a training sample with known individual group memberships, while for our problem, the group membership (genotype) of individuals (plants) in the training sample is not known.

We have considered four different versions of a normal mixture, depending on whether or not variances are homogeneous and whether or not the mixing proportions are known *a priori*. Each of these models may be useful in a given situation. With the sugar beet data, a segregation of 1:2:1 was to be expected, but apparently some loci did not fit this expectation well. It is a useful feature of the method presented here that the null hypothesis of a specific segregation rate can be tested. From the statistical point of view, parsimonious models are preferable to complicated models. This may even be true when the more complicated model is more realistic, especially when there are only limited data to

estimate the parameters. If the segregation ratio is known *a priori*, the known proportions π_i should be used rather than estimated.

Conversely, if estimation π_i is of primary interest, large samples will be needed, certainly more than $n = 46$ individuals as in our study. To obtain an idea of the necessary sample size, consider the case where the genotype of individual plants can be identified without error. Then we are in a multinomial sampling situation. The standard error of an estimate of π under multinomial sampling is $SE = \sqrt{\pi(1 - \pi)/n}$, where n is the sample size. For example, if $\pi = 0.5$, we need a sample size of $n = 100$ to achieve a standard error of 0.05. A 95% confidence limit around the estimate will then have a width of $\sim$0.2, implying a rather poor estimate. The situation becomes worse when genotypes cannot be identified without error, as in our situation.

A model with common variance tends to yield more stable results. The assumption of homogeneous variances can be tested. Especially with limited data, it often happens that one of the fitted variances is very small relative to the others. There is a danger that, *e.g.*, a group of only a few observations is erroneously recognized as a separate genotype class, with a very small variance. The likelihood of normal mixtures tends to have multiple local maxima. Also, when the variances are allowed to vary among groups, the likelihood is unbounded above (Kiefer and Wolfowitz 1956). To locate the maximum-likelihood solution, it is necessary to try several starting values. Day (1969) noted that "spurious maximizers of the likelihood, corresponding to parameter points having some component standard deviations very small relative to others, are generated by any small number of sample points grouped sufficiently close together." These observations suggest that models with homogeneous variance yield more stable results. Day (1969) recommended the use of maximum likelihood (ML) when it is known that the variances for the component densities are equal. Lynch and Walsh (1997) state that due to the practical difficulties encountered when fitting mixtures with heterogeneous variances of the components, it is often preferable to set all variances equal to each other.

Redner and Walker (1984) noted that if the component densities are poorly separated, then impractically large sample sizes might be required in order to expect even moderately precise maximum-likelihood estimates. The same authors point out that in the small sample, poor separation case for a mixture of two univariate normals, maximum-likelihood estimates should be used with extreme caution or not at all. Even with relatively good estimates based on a very large sample, poorly separated genotype groups will be hard to distinguish even by the statistical method presented here, and allocation errors will be large. Clearly, our statistical procedure cannot be expected to save a hopeless situation. It is the intermediate cases where it has greatest merit.

If OD values are bounded upward due to the measurement process, it is advisable to tune the system in such a way that measured OD values are not close to the bound. With OD values close to the bound, separation of genotypes is likely to be poor. When one $P(g = i)$ is small and the mean of the corresponding component is similar to that of another component $g = i'$, it can happen that none of the individuals is classified into the genotype class $g = i$. The likelihood of this problem occurring increases when the bulk of OD values is concentrated near a boundary. Also, nonnormality is likely to be a problem, and transformation of the data may be necessary. The measurement process may cause a large number of values to take the boundary value (*e.g.*, OD = 2). In this case, a mixture distribution with truncated component distributions (*e.g.*, truncated normals) may be useful, and fitting of these by the EM algorithm should be relatively straightforward. It is stressed again, however, that such extreme measurements should be avoided, if possible, since the information content is limited.

With our procedure, estimation of parameters and genotyping are based on the same data set. Therefore, estimates of *a posteriori* probabilities tend to be too optimistic and should be interpreted with caution. Nevertheless, using the complete data for both estimation and genotyping will be more efficient than data splitting. To have the same number of genotyped individuals, the sample size would have to be increased if estimation were to be based on independent data. This added cost is seldom justified in practice.

In summary, it is recommended to use as simple a model as possible for genotyping, preferably one with homogeneous variances and with known genotype frequencies. Our experience is that this type of model often works well and poses no numerical difficulties. Analyzing the whole data set in one step is preferable to data splitting.

We thank two referees for helpful comments.

LITERATURE CITED

Box, G. E. P., and D. R. Cox, 1964 An analysis of transformations. J. R. Stat. Soc. B **26:** 211–246.

Cox, D. R., and D. V. Hinkley, 1974 *Theoretical Statistics.* Chapman & Hall, London.

Day, N. E., 1969 Estimating the components of a mixture of normal distributions. Biometrika **56:** 463–474.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. B **39:** 1–38.

Everitt, B. S., and D. J. Hand, 1981 *Finite Mixture Distributions.* Chapman & Hall, London.

Gutierrez, R. G., R. J. Carroll, N. Wang, G.-H. Lee and B. H. Taylor, 1995 Analysis of tomato root initiation using a normal mixture distribution. Biometrics **51:** 1461–1468.

Hastie, T., and R. J. Tibshirani, 1996 Discriminant analysis by Gausian mixtures. J. R. Stat. Soc. B **58:** 155–176.

Hathaway, R. J., 1985 A constrained formulation of maximum likelihood estimation for normal mixture distributions. Ann. Stat. **13:** 795–800.
Haughton, D., 1997 Packages for estimating finite mixtures: a review. Am. Stat. **51:** 194–205.
Hochberg, Y., and A. C. Tamhane, 1987 *Multiple Comparison Procedures.* Wiley, New York.
Kiefer, J., and J. Wolfowitz, 1956 Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Stat. **27:** 887–906.
Lynch, M., and B. Walsh, 1997 *Genetics and Analysis of Quantitative Traits.* Sinauer, Sunderland, MA.
McLachlan, G. J., and K. E. Basford, 1988 *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.
McLachlan, G. J., and T. Krishnan, 1997 *The EM Algorithm and Extensions.* Wiley, New York.
Peel, D., and G. J. McLachlan, 1998 *User's Guide to EMMIX.* Version 1.0. http://www.maths.uq.edu.au/~gjm/emmix/guide.html.
Redner, R. A., and H. F. Walker, 1984 Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. **26:** 195–239.
Wolfinger, R. D., 1996 Heterogeneous variance-covariance structures for repeated measures. J. Agric. Biol. Environ. Stat. **1:** 205–230.

Communicating editor: R. G. Shaw

APPENDIX

Estimating the model: The most frequently used method of estimation is that of ML. The most popular algorithm for maximizing the likelihood of mixtures is the EM algorithm of Dempster *et al.* (1977). This algorithm, as applied to our model, is briefly sketched. For details, see, *e.g.*, McLachlan and Basford (1988) and McLachlan and Krishnan (1997). The basic idea of the EM algorithm is to regard the observed data y as incomplete. The missing data pertain to the unknown membership of the individuals to one of the genotype classes. It is convenient to introduce a random variable z_{ij} with $z_{ij} = 1$ if for the jth plant $g = i$ and $z_{ij} = 0$ otherwise. The EM algorithm regards z_{ij} ($i = 1, 2, 3; j = 1, \ldots, n$) as missing. The complete data log-likelihood is

$$\log L(\theta) = \sum_{i=1}^{G}\sum_{j=1}^{n} z_{ij}\log\{\pi_i\phi(y_j|\mu_i,\sigma_i)\}.$$

The EM algorithm alternates iteratively between two steps, the E-step (for expectation) and the M-step (for maximization). The E-step maximizes the conditional expectation of $\log L(\theta)$, given the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)'$, using the current fit $\theta^{(h)}$ for θ; *i.e.*, it maximizes

$$Q(\theta; \theta^{(h)}) = E_{\theta(h)}\{\log L(\theta)|y\}.$$

Since $\log L(\theta)$ is linear in the unobserved z_{ij}, the E-step is performed by replacing z_{ij} with its conditional expectation given y_j, using the current estimate $\theta^{(h)}$ for θ. Thus, z_{ij} is replaced by

$$E_{\theta(h)}(z_{ij}|y_j) = P(g = i|y_j; \theta^{(h)}), \qquad \text{(A1)}$$

where $P(g = i|y_j; \theta)$ is given in (2). The M-step updates $\theta^{(h)}$ with $\theta^{(h+1)}$, where $\theta^{(h+1)}$ maximizes $Q(\theta; \theta^{(h)})$. Thus, on the $(h + 1)$th step, the estimates of the mixing proportions and of θ are given by

$$\pi_i^{(h+1)} = \sum_{j=1}^{n} P(g = i|y_j;\theta^{(h)})/n,$$

$$\mu_i^{(h+1)} = \frac{\sum_{j=1}^{n} P(g = i|y_j;\theta^{(h)})y_j}{\sum_{j=1}^{n} P(g = i|y_j;\theta^{(h)})},$$

$$\sigma_i^{(h+1)} = \sqrt{\frac{\sum_{j=1}^{n} P(g = i|y_j;\theta^{(h)})(y_j - \mu_i^{(h+1)})^2}{\sum_{j=1}^{n} P(g = i|y_j;\theta^{(h)})}} \quad (i = 1, 2, 3). \qquad \text{(A2)}$$

The EM algorithm alternates repeatedly between the E-step (computation of A1) and the M-step [computation of (A2), possibly modified by fixing π_i and/or by using (A3) for the variance] until the likelihood converges (changes only by an arbitrarily small amount in successive steps).

Simplifications of the fitting procedure are effected in two cases:

1. The mixing proportions $\pi_i = P(g = i)$ are known *a priori.* In this case, π_i are fixed and the updating of π_i in the M-step can be omitted.
2. The variances of the three components are equal. In this case the common standard deviation (square root of the common variance) is estimated by

$$\sigma^{(h+1)} = \sqrt{\frac{\sum_{i=1}^{G}\sum_{j=1}^{n} P(g = i|y_j;\theta^{(h)})(y_j - \mu_i^{(h+1)})^2}{\sum_{i=1}^{G}\sum_{j=1}^{n} P(g = i|y_j;\theta^{(h)})}}. \qquad \text{(A3)}$$

It should be remarked that the likelihood for normal mixtures with heterogeneous variances tends to have multiple local maxima, so several sets of starting values need to be tried. Also, if one of the means is set equal to any of the observations and the corresponding standard deviation approaches zero, the likelihood is not bounded above, and therefore, strictly speaking, the global maximum does not exist (Kiefer and Wolfowitz 1956). Several strategies to cope with these problems (*e.g.*, that of Hathaway 1985) are implemented in the various packages for fitting normal mixtures (for a review see Haughton 1997). We used the EMMIX program (Peel and McLachlan 1998) to fit models (a) and (b). For fitting models (c) and (d), the EM algorithm was programmed using the SAS System.

Correct allocation rate: The correct allocation rate is the expected proportion of correctly classified individuals. To compute this rate, we need the classification limits y_L and y_R so that individuals are classified as follows:

Classification	Condition for OD value y
$g = 1$	$y < y_L$
$g = 2$	$y_L \le y \le y_R$
$g = 3$	$y_R < y$

The classification limit y_L is the point of intersection between $\pi_1\phi(y_j|\mu_1, \sigma_1) = \pi_2\phi(y_j|\mu_2, \sigma_2)$ for $\mu_1 < y_L < \mu_2$.

The classification limit y_R is the point of intersection between $\pi_2\phi(y_j|\mu_2, \sigma_2) = \pi_3\phi(y_j|\mu_3, \sigma_3)$ for $\mu_2 < y_R < \mu_3$ (see Figure 3). For homogeneous (heterogeneous) σ_i, y_L and y_R are the solution of a linear (quadratic) equation in y_L and y_R, respectively. For example, when σ_i is homogeneous,

$$y_L = \frac{\sigma^2 \log(\pi_1/\pi_2)}{\mu_2 - \mu_1} + \frac{\mu_2 + \mu_1}{2} \tag{A4}$$

$$y_R = \frac{\sigma^2 \log(\pi_2/\pi_3)}{\mu_3 - \mu_2} + \frac{\mu_3 + \mu_2}{2}. \tag{A5}$$

Let $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal and CAR_i the correct allocation rate for the ith component. Then

$$\begin{aligned} CAR_1 &= \Phi\left(\frac{y_L - \mu_1}{\sigma_1}\right) \\ CAR_2 &= \Phi\left(\frac{y_R - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{y_L - \mu_2}{\sigma_2}\right) \\ CAR_3 &= 1 - \Phi\left(\frac{y_R - \mu_3}{\sigma_3}\right). \end{aligned} \tag{A6}$$

The overall correct classification rate is

$$CAR = \sum_{i=1}^{g} \pi_i CAR_i. \tag{A7}$$

We estimate CAR by plugging in sample estimates for parameters. This approach provides a rough assessment of the true CAR.