Theor Appl Genet (2001) 103:462–468

Hans-Peter Piepho

Exploiting quantitative information in the analysis of dominant markers

Received: 2 May 2000 / Accepted: 6 December 2000

Abstract Dominant genetic markers such as AFLPs and RAPDs are usually analyzed based on the presence or absence of a band on an electrophoretic gel. This type of analysis does not allow a distinction among dominant homozygotes and heterozygotes. Such a distinction is possible based on the quantitative measurement of band intensities. In the present paper, we consider the problem of analyzing dominant markers based on band-intensity data. The basic step for mapping a marker is to assess its recombination frequency with other markers. Ordering markers on a map can then be done using a number of standard procedures. For this reason estimation of the recombination frequency is the main focus of the present paper. The method is demonstrated for the case of an F_2 population. By simulation we investigate its accuracy and compare it to the standard estimation based on dominant scoring for band presence/absence. There are a number of potential applications. For example, the map may be used to locate quantitative trait loci (QTLs), applying standard procedures modified to account for uncertainty of the marker genotype. Moreover, map information can be used to determine the most likely genotype at a marker, given its band intensity and the band intensities at flanking markers.

Keywords QTL analysis · AFLP · Genotyping · Bayes theorem · Two-point analysis · Recombination fraction · Linkage mapping

Introduction

Consider an F_2-population derived from a single parental cross. We will look at the inheritance of two linked loci. Let M_1 and m_1 be the alleles at the first locus, while the alleles at the second locus are M_2 and m_2. Our objective is to estimate the recombination fraction between the two marker loci, which is denoted here as ρ. The allele M_k (k=1, 2) produces a band on an electrophoretic gel. It is dominant with respect to the null allele m_k. Thus, absence of a band is detected only for the genotype m_km_k. While genotypes M_kM_k and M_km_k both produce a band, the band intensity is expected to be larger for the former. Thus, with a quantitative measurement of band intensities, M_kM_k and M_km_k can be distinguished, provided measurement errors are not too large.

The purpose of the present paper is to demonstrate how the recombination fraction can be estimated if band intensities are available. We will use a normal mixture model for this purpose and estimate parameters by the expectation-maximization (EM) algorithm (McLachlan and Krishnan 1997). The method is exemplified using a real data set. We discuss potential applications, i.e. construction of marker maps, quantitative trait loci (QTLs) mapping, and codominant scoring of AFLP markers.

Communicated by H.C. Becker

H.-P. Piepho
Institut für Nutzpflanzenkunde, Universität Kassel,
Steinstraße 19, 37213 Witzenhausen, Germany
e-mail: piepho@wiz.uni-kassel.de

Materials and methods

Model

Assume that for the r-th plant (r=1, ..., n), we have a quantitative measurement at each marker position k (k=1, 2). Denote this measurement as y_{kr}. Based on y_{kr} we want to estimate the recombination fraction ρ between the two markers. An F_1-individual M_1M_2/m_1m_2 produces nine distinguishable genotypes (Tab. 2). Let g_k (k=1, 2) be a random variable with g_k=1 for marker genotype m_km_k, g_k=2 for M_km_k, and g_k=3 for M_kM_k. Denote the joint probability that $g_1=i$ and $g_2=j$ (i,j=1, 2, 3) as

$$\pi_{ij}(\rho)=\mathrm{P}(g_1=i, g_2=j). \quad (1)$$

The probability π_{ij} depends on ρ, the recombination frequency between the marker loci, as shown in Table 1 (Ott 1991, 1993; Weber and Wricke 1994). Measurements of band intensitities for marker genotypes M_kM_k and M_km_k will be subject to measurement errors. Moreover, any device for measuring band intensity is likely to give a non-zero measurement for m_km_k individuals, which may be regarded as background noise (Piepho and Koch 2000). This measurement, too, will show some measurement error. Thus, con-

Table 1 Marker genotypes in the F_2 derived from the cross $M_1M_1M_2M_2 \times m_1m_1m_2m_2$, where the two loci have the recombination fraction ρ

Genotype		i	j	$\pi_{ij}(\rho)$	$S_{ij}(\rho)$[a]	Coefficient of the polynomial obtained by multiplying $S_{ij}(\rho)$ with $\rho(1-\rho)(1-2\rho+2\rho^2)$			
Marker 1	Marker 2					ρ^0	ρ^1	ρ^2	ρ^3
M_1M_1	M_2M_2	3	3	$(1-\rho)^2/4$	$-2/(1-\rho)$	0	–2	4	–4
M_1M_1	M_2m_2	3	2	$\rho(1-\rho)/2$	$(1-2\rho)/[\rho(1-\rho)]$	1	–4	6	–4
M_1M_1	m_2m_2	3	1	$\rho^2/4$	$2/\rho$	2	–6	8	–4
M_1m_1	M_2M_2	2	3	$\rho(1-\rho)/2$	$(1-2\rho)/[\rho(1-\rho)]$	1	–4	6	–4
M_1m_1	M_2m_2	2	2	$(1-2\rho+2\rho^2)/2$	$2(2\rho-1)/(1-2\rho+2\rho^2)$	0	–2	6	–4
M_1m_1	m_2m_2	2	1	$\rho(1-\rho)/2$	$(1-2\rho)/[\rho(1-\rho)]$	1	–4	6	–4
m_1m_1	M_2M_2	1	3	$\rho^2/4$	$2/\rho$	2	–6	8	–4
m_1m_1	M_2m_2	1	2	$\rho(1-\rho)/2$	$(1-2\rho)/[\rho(1-\rho)]$	1	–4	6	–4
m_1m_1	m_2m_2	1	1	$(1-\rho)^2/4$	$-2/(1-\rho)$	0	–2	4	–4

[a] $S_{ij}(\rho) = \frac{1}{\pi_{ij}(\rho)} \frac{\partial \pi_{ij}(\rho)}{\partial \rho}$

ditionally on the marker genotype s (s=1, .., 3 for marker genotypes *mm*, *Mm* and *MM*, respectively), band intensities at the k-th marker are assumed to follow a normal distribution with a mean μ_{ks} and a variance σ^2_{ks}. Let

$$\phi(y_{kr}|\mu_{ks}, \sigma^2_{ks}) \tag{2}$$

be the normal density for y_{kr}, the quantitative measurement on the k-th marker for the r-th plant (r=1, .., n; for simplicity, no distinction is made in notation between a random variable and its realized values). The density has a mean μ_{ks} and a variance σ^2_{ks}. Conditionally on the marker genotype, band intensities at different markers will be assumed to be independent. This is a convenient, though approximate, assumption, since bands on the same lane are expected to show some correlation. If lanes are corrected based on information from monomorphic bands, the independence assumption is often tenable (Piepho and Koch 2000). There are a number of software packages that use this type of correction, e.g. the AFLP-Quantar package by KeyGene (www.keygene.com). Assuming independence, the marginal distribution of $y_r=(y_{1r}, y_{2r})$, the observation vector for the r-th plant, is a mixture of nine bivariate normal distributions:

$$f(y_r|\theta) = \sum_{i=1}^{3}\sum_{j=1}^{3} \pi_{ij}(\rho)\phi(y_{1r}|\mu_{1i}, \sigma^2_{1i})\phi\left(y_{2r}\middle|\mu_{2j}, \sigma^2_{2j}\right), \tag{3}$$

where $\theta=[\rho, (\mu_{ks}), (\sigma^2_{ks})]$ (markers k=1, 2; marker genotypes s=1, ..., 3). The model can be simplified by assuming variance homogeneity at different levels, e.g.,

$$\sigma^2_{ks}=\sigma^2_k, \text{ or} \tag{4}$$

$$\sigma^2_{ks}=\sigma^2. \tag{5}$$

In a model with a single marker, assuming variance homogeneity at the marker (as in equation 4) has been found to be appropriate in many cases and to yield more stable results than a model with a separate variance component for each marker genotype (Piepho and Koch 2000). Assuming variance homogeneity across markers (equation 5) may be unrealistic though, since the type and quantity of product measured at a band position varies among markers. For these reasons, we will henceforth use model (4).

Two-point analysis by the EM algorithm

The method

Let z_{ijr} be a random variable with z_{ijr}=1 if $g_1=i$ and $g_2=j$ and z_{ijr}=0 otherwise. $z_r=(z_{11r}, ..., z_{33r})$ follows a multinomial distribution with a constant 1 and cell probabilities of $\pi_{ij}(\rho)$. The problem in practice is that z_r is not observed. Estimation would be straightforward if both y_r and z_r were fully observed. This is exploited by the expectation-maxmization (EM) algorithm (McLachlan and Krishnan 1997). Essentially, the EM algorithm for fitting (3) iterates between two steps, one that imputes the missing data (z_r) and one that estimates the parameters from the completed data. Using the terminology common with the EM algorithm, we may say that $x_r=(y_r, z_r)$ is the complete data vector and y_r is the incomplete data vector. In the sequel we describe the technical details of the algorithm. Readers not interested in these technicalities may proceed directly to the example.

The likelihood for the complete data is

$$\log L(\theta) = \sum_{r=1}^{n}\sum_{i=1}^{3}\sum_{j=1}^{3} z_{ijr} \log\left[\pi_{ij}(\rho)\phi(y_{1r}|\mu_{1i}, \sigma^2_1)\phi\left(y_{2r}\middle|\mu_{2j}, \sigma^2_2\right)\right]. \tag{6}$$

The conditional expectation of the complete-data log likelihood, given the observed data y_r, using the current estimates for the parameters $\theta^{(h)}$, may be expressed as

$$Q(\theta;\theta^{(h)}) = E_{\theta^{(h)}}[\log L(\theta^{(h)})|y_r]. \tag{7}$$

Since $Q(\theta;\theta^{(h)})$ is linear in z_{ijr}, $Q(\theta;\theta^{(h)})$ is computed by replacing z_{ijr} by its expectation, given y_r, in the complete-data log-likelihood evaluated at $\theta;\theta^{(h)}$; i.e. z_{ijr} is replaced by

$$w^{(h)}_{ijr} = \frac{\pi_{ij}(\rho)\phi\left(y_{1r}\middle|\mu^{(h)}_{1i}, \sigma^{2(h)}_1\right)\phi\left(y_{2r}\middle|\mu^{(h)}_{2j}, \sigma^{2(h)}_2\right)}{\sum_{i=1}^{3}\sum_{j=1}^{3} \pi_{ij}(\rho)\phi\left(y_{1r}\middle|\mu^{(h)}_{1i}, \sigma^{2(h)}_1\right)\phi\left(y_{2r}\middle|\mu^{(h)}_{2j}, \sigma^{2(h)}_2\right)}. \tag{8}$$

The M-step maximizes $Q[\theta;\theta^{(h)}]$ with respect to the parameters, while the E-step updates $w^{(h)}_{ijr}$ based on the current parameter estimates. In the M-step, we can exploit the fact that $\rho^{(h)}$ and $[\mu^{(h)}_{ks}, \sigma^{2(h)}_k]$ are disjoint in $Q(\theta;\theta^{(h)})$. The estimating equations for $[\mu^{(h)}_{ks}, \sigma^{2(h)}_k]$ have explicit solutions, while the equation for $\rho^{(h)}$ is a third-degree polynomial, which may be solved numerically, e.g. using the POLYROOT function of SAS/IML, or by explicit formulae (Abramowitz and Stegun 1972). The EM-algorithm proceeds as follows:

(1) Initial step: choose starting values for θ.
(2) E-step: update the weights $w^{(h)}_{ijr}$.
(3) M-Step: estimates for $[\mu^{(h)}_{ks}, \sigma^{2(h)}_k]$ (k=1, 2; s=1, 2, 3) are updated as follows:

$$\left.\begin{aligned} \mu^{(h+1)}_{1s} &= \frac{\sum_{r=1}^{n}\sum_{j=1}^{3} w^{(h)}_{sjr} y_{1r}}{\sum_{r=1}^{n}\sum_{j=1}^{3} w^{(h)}_{sjr}}; \quad \mu^{(h)}_{2s} = \frac{\sum_{r=1}^{n}\sum_{i=1}^{3} w^{(h)}_{isr} y_{2r}}{\sum_{r=1}^{n}\sum_{i=1}^{3} w^{(h)}_{isr}} \\ \sigma^{2(h)}_1 &= \frac{\sum_{r=1}^{n}\sum_{i=1}^{3}\sum_{j=1}^{3} w^{(h)}_{ijr}\left(y_{1r}-\mu^{(h)}_{1i}\right)^2}{n}; \\ \sigma^{2(h)}_2 &= \frac{\sum_{r=1}^{n}\sum_{i=1}^{3}\sum_{j=1}^{3} w^{(h)}_{ijr}\left(y_{2r}-\mu^{(h)}_{2j}\right)^2}{n}. \end{aligned}\right\} \tag{9}$$

Update $\rho^{(h)}$ by the non-complex root of

$$\frac{\partial Q(\theta;\theta^{(h)})}{\partial \rho} = \sum_{r=1}^{n}\sum_{i=1}^{3}\sum_{j=1}^{3} w^{(h)}_{ijr} S_{ij}(\rho) = 0, \tag{10}$$

where

$$S_{ij}(\rho) = \frac{1}{\pi_{ij}(\rho)} \frac{\partial \pi_{ij}(\rho)}{\partial \rho},$$

Table 2 Band intensity values for two markers from the sugar beet experiment described in Piepho and Koch (2000). The 46 individuals appear in the same order for both markers

Marker 1									
0.504	1.022	1.001	0.250	0.730	0.520	0.840	0.442	1.083	1.075
1.118	0.910	0.334	1.028	0.722	0.622	0.798	0.554	0.340	0.937
0.395	0.973	0.811	0.719	1.019	0.869	0.640	0.441	0.943	0.795
0.747	0.470	0.744	0.878	0.882	0.707	0.862	1.238	0.291	0.278
0.255	0.323	0.262	0.801	0.870	0.350				
Marker 2									
0.534	1.182	1.041	0.260	0.290	0.270	1.050	0.272	1.183	1.155
1.008	0.900	1.104	1.168	0.952	0.222	0.948	0.894	0.810	1.087
0.225	1.163	1.031	0.839	1.089	0.729	0.840	0.141	1.123	0.935
0.887	1.150	0.904	1.158	1.142	0.247	0.892	1.238	0.951	0.918
0.795	0.923	0.822	1.031	0.840	0.250				

that maximizes $Q(\theta;\theta^{(h)})$. In our experience, this equation has exclusively had one non-complex root and one conjugate pair of complex roots. If the solution is larger than 0.5, set $\rho^{(h+1)}=0.5$. If the solution is smaller than 0, set $\rho^{(h+1)}=0.0$.

Iterate steps (2) and (3) until convergence.

Results and discussion

Example

To exemplify the method, we use band intensity data for two linked AFLP markers from a sugar beet data set described in detail by Piepho and Koch (2000). Due to the measurement device, band intensities ranged from about 0 to about 2. The data are displayed in Table 2. Application of the EM algorithm to these data yields the following parameter estimates: $\hat{\mu}_{11}=0.369$, $\hat{\mu}_{12}=0.770$, $\hat{\mu}_{13}=1.012$, $\hat{\sigma}_1=0.0983$, $\hat{\mu}_{21}=0.271$, $\hat{\mu}_{22}=0.898$, $\hat{\mu}_{23}=1.128$, $\hat{\sigma}_2=0.0780$ and $\hat{\rho}=0.191$. Thus, the two markers were only loosely linked.

Simulation experiment

We will compare the estimator for quantitative marker data (quantitative method) with the estimator for band presence/absence data (qualitative method). Under the assumed mixture model, band intensities follow a normal distribution and the normal components show some degree of overlap. Thus, even if we assume that band intensity is assessed correctly by eye, the classification of a band as present or absent will be subject to misclassification error. The error increases with increasing variance of the mixing components and with decreasing distance between means. To mimic the assessment of band presence or absence in practice, we will assume that the rater has full knowledge of the underlying distribution and determines the genotype using the posterior genotype probabilities derived from the Bayes theorem as described in Piepho and Koch (2000). According to this decision rule, a band at the first marker will be rated as absent if

$$y_{1r} < \frac{\sigma_1^2 \log[(\pi_{11}+\pi_{12}+\pi_{13})/(\pi_{21}+\pi_{22}+\pi_{23})]}{\mu_{12}-\mu_{11}} + \frac{\mu_{12}+\mu_{11}}{2}. \quad (11)$$

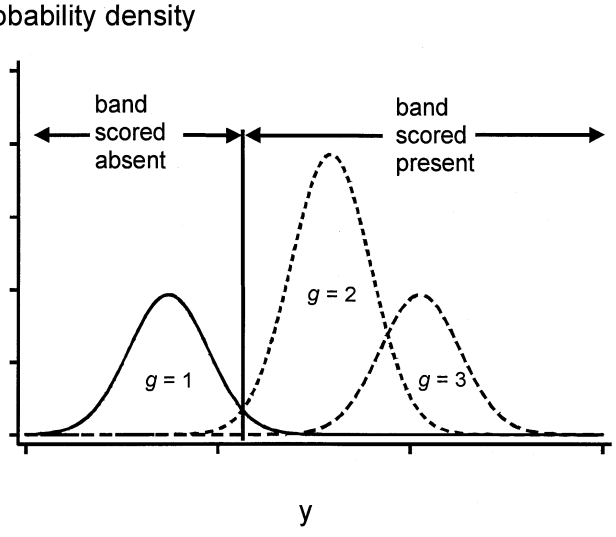

Fig. 1 Illustration of the decision rule for the dominant scoring procedure used in simulation

This threshold corresponds the solution for y_{1r} in

$$(\pi_{11}+\pi_{12}+\pi_{13})\phi(y_{1r}|\mu_{11},\sigma_1^2) = (\pi_{21}+\pi_{22}+\pi_{23})\phi(y_{1r}|\mu_{12},\sigma_1^2), \quad (12)$$

i.e. the point of intersection for the two normal distributions of m_1m_1 and m_1M_1, weighted by the marginal *a priori* probabilities of these two genotypes. The equivalent rule for the second marker is

$$y_{2r} < \frac{\sigma_2^2 \log[(\pi_{11}+\pi_{21}+\pi_{31})/(\pi_{12}+\pi_{22}+\pi_{23})]}{\mu_{22}-\mu_{21}} + \frac{\mu_{22}+\mu_{21}}{2}. \quad (13)$$

The decision rule is illustrated in Fig. 1, assuming that the rating of band absence/presence following this decision rule is optimistic. In reality, the rater will not have full knowledge of the true distribution, and the misclassification rate will be larger than under the Bayes rule in (12) and (13). Thus, our comparison favors the qualitative method. Moreover, markers were assumed to be in coupling phase, which is expected to be a much more favorable setting for the qualitative method than repulsion linkage (Liu et al. 1998, p 193). Under the qualitative method, the recombination fraction will be estimated us-

ing standard procedures (Weir 1998, p 236). This assumes that all individuals are classified correctly, which is not the case under the assumed model. Thus, some bias and variance inflation is expected compared to a situation without a misclassification error.

To compare the qualitative and quantitative methods, we simulated 1,000 data sets and computed the mean squared error (MSE), the root mean squared error (RMSE), the standard deviation (SD), and the bias (BIAS) for both estimators of ρ. In simulations, the following parameters were varied: sample size, n (50, 200, 500), means (μ_{11}, μ_{12}, μ_{13}), [(0.5, 1.0, 1.5), (0.5, 1.2, 1.5), (0.5, 1.35, 1.5), (0.5, 1.45, 1.5)], standard deviations, σ_1 (0.05, 0.10), and recombination fraction, ρ (0.01, 0.05, 0.20). The same means and variances were assumed for both markers. The choices for means and variances are based on experience with AFLP band-intensity data (Piepho and Koch 2000).

The results are reported in Tables 3, 4 and 5. Throughout, the quantitative method has the more favorable values for MSE, RMSE, SD and BIAS, or the two methods are comparable. The better-separated are the normal components for $M_k m_k$ and $M_k M_k$, i.e. the farther apart the means and the smaller the variance, the higher the gain. When the means for $M_k m_k$ and $M_k M_k$ are very close, the methods are virtually identical in performance. Thus, using band intensity data promises to often provide a gain in accuracy compared to presence/absence data. In the worst case, the gain is small or negligible. It appears that nothing can be lost by trying to exploit quantitative information.

Some fields of application

In what follows, I will consider potential fields of application for the proposed method, i.e. map construction, QTL mapping and genotyping of AFLP markers.

Map construction

The simulations have shown that it is worthwhile to exploit band intensity information. The suggested method allows a co-dominant analysis, which accounts for uncertainty regarding the marker genotype in case of measurement errors. Such errors are quite common in practice (Piepho and Koch 2000). Of course, co-dominant markers, which allow an unequivocal distinction among heterozygotes and dominant homozygotes, are preferable due to higher information content; but dominant markers such as AFLPs are often used for economical reasons. Co-dominant scoring of such markers based on intensity data allows more information to be extracted than an analysis based on the presence or absence of a band. Once all pairwise recombination fractions have been estimated for a set of markers, one can use one of a number of standard methods for genetic map-construction (Weir 1996, p 241; Liu 1998), including the least-squares approach suggested by Jensen and Jørgensen (1975; and see Stam 1993). Also, the method of Lander and Green (1987) and Lander et al. (1987), which does not utilize pairwise recombination estimates, is formulated generally enough to allow quantitative data to be handled. Since their method is markedly different from the present two-point approach, I will not elaborate this option here.

Table 3 Mean square error (MSE), root mean square error (RMSE), standard deviation (SD) and bias (BIAS) for estimates of the recombination fraction based on presence/absence data (qualitative) and on band intensity data (quantitative). Results based on 1,000 simulations for each setting. n=50

Parameters[a]					MSE×10,000		RMSE×100		SD×100		BIAS×100	
μ_{11}	μ_{12}	μ_{13}	σ_1	ρ	Qual.	Quant.	Qual.	Quant.	Qual.	Quant.	Qual.	Quant.
0.5	1.0	1.5	0.05	0.01	2.30	1.02	1.52	1.01	1.51	1.01	0.083	0.045
				0.05	10.07	5.33	3.17	2.31	3.17	2.31	0.058	0.107
				0.20	45.34	22.09	6.73	4.70	6.73	4.69	0.290	0.295
			0.10	0.01	4.52	1.05	2.13	1.03	1.92	1.03	0.907	–0.004
				0.05	12.26	5.54	3.50	2.35	3.41	2.35	0.790	–0.115
				0.20	50.39	22.95	7.10	4.79	7.04	4.79	0.911	0.188
0.5	1.2	1.5	0.05	0.01	2.10	0.98	1.45	0.99	1.45	0.99	–0.010	0.003
				0.05	9.89	4.85	3.14	2.20	3.14	2.20	–0.108	–0.038
				0.20	41.30	20.97	6.43	4.58	6.42	4.58	–0.170	–0.003
			0.10	0.01	2.11	1.53	1.45	1.24	1.45	1.23	0.070	–0.028
				0.05	10.47	7.23	3.24	2.69	3.23	2.69	0.188	0.013
				0.20	45.94	29.42	6.78	5.42	6.77	5.42	0.270	–0.029
0.5	1.35	1.5	0.05	0.01	2.06	1.63	1.44	1.27	1.44	1.27	0.027	–0.022
				0.05	9.90	7.51	3.15	2.74	3.14	2.74	0.104	–0.025
				0.20	44.45	27.03	6.67	5.20	6.66	5.20	0.223	–0.023
			0.10	0.01	1.88	1.85	1.37	1.36	1.37	1.36	–0.014	–0.029
				0.05	9.53	9.32	3.08	3.05	3.08	3.05	0.133	0.068
				0.20	42.09	38.79	6.49	6.23	6.49	6.22	–0.143	–0.350
0.5	1.45	1.5	0.05	0.01	2.17	2.17	1.47	1.47	1.47	1.47	0.061	0.055
				0.05	9.83	9.64	3.14	3.11	3.13	3.11	0.084	0.021
				0.20	45.90	42.55	6.77	6.52	6.76	6.52	0.469	0.188
			0.10	0.01	1.97	1.94	1.40	1.39	1.40	1.39	–0.055	–0.060
				0.05	10.89	10.97	3.30	3.31	3.30	3.31	0.065	0.033
				0.20	43.70	40.88	6.61	6.39	6.61	6.39	0.227	–0.110

[a] Means and variances are the same for both markers

Table 4 Mean square error (MSE), root mean square error (RMSE), standard deviation (SD) and bias (BIAS) for estimates of the recombination fraction based on presence/absence data (qualitative) and on band intensity data (quantitative). Results based on 1,000 simulations for each setting. n=200

Parameters[a]					MSE×10,000		RMSE×100		SD×100		BIAS×100	
μ_{11}	μ_{12}	μ_{13}	σ_1	ρ	Qual.	Quant.	Qual.	Quant.	Qual.	Quant.	Qual.	Quant.
0.5	1.0	1.5	0.05	0.01	0.53	0.26	0.72	0.51	0.72	0.51	0.004	–0.013
				0.05	2.43	1.20	1.56	1.09	1.56	1.09	0.023	0.013
				0.20	10.51	4.96	3.24	2.23	3.24	2.23	0.032	0.055
			0.10	0.01	1.65	0.27	1.28	0.52	0.93	0.52	0.883	–0.018
				0.05	3.46	1.32	1.86	1.15	1.71	1.15	0.730	–0.018
				0.20	10.65	5.25	3.26	2.29	3.21	2.29	0.583	0.006
0.5	1.2	1.5	0.05	0.01	0.51	0.27	0.71	0.52	0.71	0.52	0.006	–0.004
				0.05	2.41	1.20	1.55	1.09	1.55	1.09	–0.056	0.002
				0.20	10.31	5.12	3.21	2.26	3.21	2.26	0.136	0.090
			0.10	0.01	0.53	0.43	0.73	0.65	0.73	0.65	0.034	–0.004
				0.05	2.52	1.77	1.59	1.33	1.59	1.33	0.027	–0.034
				0.20	10.13	6.13	3.18	2.48	3.18	2.47	–0.125	–0.146
0.5	1.35	1.5	0.05	0.01	0.50	0.41	0.71	0.64	0.71	0.64	–0.016	–0.036
				0.05	2.45	1.74	1.57	1.32	1.57	1.32	0.035	0.053
				0.20	10.54	6.29	3.25	2.51	3.24	2.51	0.209	0.055
			0.10	0.01	0.55	0.54	0.74	0.73	0.74	0.73	0.022	0.014
				0.05	2.59	2.44	1.61	1.56	1.61	1.56	–0.017	–0.051
				0.20	10.03	8.86	3.17	2.98	3.16	2.98	0.125	–0.018
0.5	1.45	1.5	0.05	0.01	0.48	0.48	0.69	0.69	0.69	0.69	–0.007	–0.011
				0.05	2.53	2.52	1.59	1.59	1.59	1.58	0.112	0.091
				0.20	10.06	9.75	3.17	3.12	3.17	3.12	0.055	–0.032
			0.10	0.01	0.49	0.49	0.70	0.70	0.70	0.70	0.000	–0.002
				0.05	2.69	2.68	1.64	1.64	1.64	1.64	–0.037	–0.048
				0.20	10.10	9.97	3.18	3.16	3.18	3.16	0.097	0.050

[a] Means and variances are the same for both markers

Table 5 Mean square error (MSE), root mean square error (RMSE), standard deviation (SD) and bias (BIAS) for estimates of the recombination fraction based on presence/absence data (qualitative) and on band intensity data (quantitative). Results based on 1,000 simulations for each setting. n=500

Parameters[a]					MSE×10,000		RMSE×100		SD×100		BIAS×100	
μ_{11}	μ_{12}	μ_{13}	σ_1	ρ	Qual.	Quant.	Qual.	Quant.	Qual.	Quant.	Qual.	Quant.
0.5	1.0	1.5	0.05	0.01	0.20	0.10	0.45	0.31	0.45	0.31	0.019	0.011
				0.05	1.08	0.53	1.04	0.73	1.04	0.72	0.020	0.039
				0.20	4.44	2.24	2.11	1.50	2.10	1.49	0.103	0.061
			0.10	0.01	1.08	0.11	1.04	0.33	0.62	0.33	0.833	–0.015
				0.05	1.75	0.50	1.32	0.71	1.09	0.71	0.745	–0.020
				0.20	4.52	2.14	2.13	1.46	2.06	1.46	0.527	0.017
0.5	1.2	1.5	0.05	0.01	0.19	0.10	0.44	0.31	0.44	0.31	–0.011	0.001
				0.05	0.98	0.47	0.99	0.69	0.99	0.69	–0.069	–0.040
				0.20	3.83	1.96	1.96	1.40	1.96	1.40	0.001	–0.026
			0.10	0.01	0.20	0.16	0.44	0.39	0.44	0.39	0.046	0.006
				0.05	0.99	0.68	0.99	0.82	0.99	0.82	–0.030	–0.049
				0.20	4.18	2.59	2.04	1.61	2.04	1.61	0.039	–0.006
0.5	1.35	1.5	0.05	0.01	0.20	0.16	0.45	0.40	0.45	0.40	0.023	0.016
				0.05	1.02	0.70	1.01	0.84	1.01	0.84	–0.027	–0.025
				0.20	4.19	2.59	2.05	1.61	2.05	1.61	0.050	0.015
			0.10	0.01	0.22	0.22	0.47	0.46	0.47	0.46	0.002	–0.001
				0.05	1.05	1.01	1.03	1.01	1.03	1.01	–0.004	–0.005
				0.20	4.11	3.82	2.03	1.95	2.03	1.95	–0.044	–0.083
0.5	1.45	1.5	0.05	0.01	0.19	0.18	0.43	0.43	0.43	0.43	0.004	0.003
				0.05	1.03	1.03	1.01	1.01	1.01	1.01	0.024	0.017
				0.20	4.43	4.23	2.10	2.06	2.10	2.06	0.131	0.093
			0.10	0.01	0.21	0.21	0.45	0.45	0.45	0.45	0.026	0.024
				0.05	0.99	0.99	0.99	0.99	0.99	0.99	0.030	0.026
				0.20	3.90	3.90	1.97	1.97	1.97	1.97	–0.050	–0.074

[a] Means and variances are the same for both markers

Whatever method is used for map construction, the resulting map will be more accurate with a quantitative analysis than with the analogous qualitative analysis, since the recombination fractions are estimated more accurately with the former.

Mapping QTLs

Genetic maps are the basis for QTL mapping. Maps based on quantitative markers can be used for QTL mapping with some modification of standard procedures. The key quantities for QTL mapping procedures such as interval mapping (IM; Lander and Bostein 1989) or com-

posite interval mapping (CIM; Zeng 1994) are the conditional genotype probabilities at a putative QTL, given the flanking markers (Lynch and Walsh 1998). These probabilities are needed for both the maximum likelihood and the least squares method of estimation (Knapp et al. 1990; Haley and Knott 1992). Procedures such as IM and CIM were developed assuming known marker genotypes. In the case of quantitative markers, the marker genotypes are known only with uncertainty, since we only have the band intensities, y_r, but not the marker genotypes. Thus, to map QTLs using quantitative markers, we simply need to replace the conditional QTL genotype probabilities, given the marker genotypes, by the conditional QTL genotype probabilities, given the band intensities, y_r. All other computational steps remain essentially unaltered.

Appealing to the Bayes Theorem (Cox and Hinkley 1974), we may compute the *a posteriori* probabilities of marker genotypes for the *r*-th individual, given the band intensitities y_r, as follows:

$$\mathrm{P}(g_1 = i, g_2 = j|y_r;\theta) = \frac{\pi_{ij}(\rho)\phi(y_{1r}|\mu_{1i},\sigma^2_{1i})\phi\left(y_{2r}|\mu_{2j},\sigma^2_{2j}\right)}{f(y_r|\theta)} \tag{14}$$

Let Q be a random variable denoting the QTL genotype at the putative QTL. For an F_2, Q has three possible realized values. Denote the conditional probability that Q=h, given the flanking markers as P(Q=h|g_1=i, g_2=j) (i, j, h=1, 2, 3). These probabilities may be computed assuming either complete interference (Knapp et al. 1990), or absence of interference (Haley and Knott 1992), as in standard QTL mapping by IM or CIM. Then, the conditional probability of the QTL genotype, given y_r is,

$$\mathrm{P}(Q = h|y_r;\theta) = \sum_{i,j}\mathrm{P}(Q = h|g_1 = i, g_2 = j)\mathrm{P}(g_1 = i, g_2 = j|y_r;\theta) \tag{15}$$

The only necessary modification of standard QTL mapping procedures is to replace P(Q=h|g_1=i, g_2=j), which is used in case of known marker genotypes, by P(Q=h|y_r; θ) in (15). To employ this approach in practice, a good estimate of θ as well as a measurement of y_r for each individual of the population to be mapped are required.

Genotyping AFLP markers

There is currently much interest in co-dominant scoring of AFLP markers, and some specific software is available for this purpose, e.g. AFLP-Quantar by Keygene (www.keygene.com). Recently, Piepho and Koch (2000) have presented a mixture approach for co-dominant scoring, which fits a normal mixture to the band intensities and then computes *a posteriori* probabilities of genotypes, given the band intensities. This approach is map-independent, since it only uses information on the marker to be scored. If a map is available, information from flanking markers can be used to improve scoring. Provided that both the markers to be scored as well as the flanking markers are AFLPs with measured band intensities, Bayes Theorem may be applied in a straightforward manner to compute conditional probabilities for the marker to be scored.

Consider an F_2 population derived from an F_1 plant $M_1M_2M_3/m_1m_2m_3$. Let ρ_1 and ρ_2 denote the recombination factions between markers 1 and 2 and between markers 2 and 3, respectively. Assume that marker 2 is flanked by marker 1 to the left and marker 3 to the right. Further, assume that for the *r*-th plant we have a quantitative measurement at each marker position k (k=1, 2, 3). Denote this measurement as y_{kr}. Based on y_r=(y_{1r}, y_{2r}, y_{3r}) we want to infer the most-likely genotype at k=2, exploiting the information both at the marker itself and at the flanking markers k=1 and k=3. Let g_k be a discrete random variable for the k-th marker locus, which denotes the genotype at that locus. The random variable can take values 1, 2 and 3 corresponding to the three possible genotypes at the locus, as described in Table 1. Denote the *a priori* genotype probabilities as

$$\pi_{ijh}(\rho_1, \rho_2)=\mathrm{P}(g_1=i, g_2=j, g_3=h)\ (i, j, h=1, 2, 3). \tag{16}$$

As indicated by the notation, these probabilities depend on the recombination fractions ρ_1 and ρ_2. Explicit formulae may be found, e.g., in van Ooijen (1992). Conditionally on the marker genotype g_k=s, band intensities at the k-th marker are assumed to follow a normal distribution with a mean μ_{ks} and a variance σ^2_k. Thus, the marginal distribution of y_r=(y_{1r}, y_{2r}, y_{3r}) is a mixture of 27 trivariate normal distributions:

$$\begin{aligned} f(y_r|\theta) = &\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{h=1}^{3}\pi_{ij}(\rho_1,\rho_2)\phi(y_{1r}|\mu_{1i},\sigma^2_1) \\ &\cdot\phi\left(y_{2r}|\mu_{2j},\sigma^2_2\right)\phi(y_{3r}|\mu_{3h},\sigma^2_3), \end{aligned} \tag{17}$$

where

$$\theta=[\rho_1, \rho_2, (\mu_{ks}), (\sigma^2_k)] \quad (k, s=1, ..., 3). \tag{18}$$

According to the Bayes Theorem (Cox and Hinkley 1974) the *posterior* probability of the genotype class membership of an individual at marker locus k=2, given its phenotypic value at the three markers (y_r), is

$$\begin{aligned} \tau_{jr} = \mathrm{P}(g_2 = j|y_r;\theta) &= \frac{\mathrm{P}(g_2 = j)f(y_r;\theta|g_2 = j)}{f(y_r|\theta)} \\ &= \frac{\sum_{i=1}^{3}\sum_{h=1}^{3}\pi_{ij}(\rho_1,\rho_2)\phi(y_{1r}|\mu_{1i},\sigma^2_1)\phi\left(y_{2r}|\mu_{2j},\sigma^2_2\right)\phi(y_{3r}|\mu_{3h},\sigma^2_3)}{f(y_r|\theta)}. \end{aligned} \tag{19}$$

This probability can be used for finding the most likely genotypes of AFLP markers, given band intensities at the marker in question and the two flanking markers. Again, practical use requires good estimates of the parameters θ. For example, means and variances of the normal mixtures can be estimated using the method described in Piepho and Koch (2000), while recombination fractions are estimated using the two-point method in the present paper.

To evaluate the merit of co-dominant scoring, we may compute the correct allocation rate (*CAR*), i.e. the probability that a randomly drawn individual is classified

correctly with respect to g_2. Following Basford and McLachlan (1985) and McLachlan and Basford (1988) we can estimate the *CAR* by

$$T = \sum_{r=1}^{n} \max_{j} \hat{\tau}_{jr} / n \qquad (20)$$

where $\hat{\tau}_{jr}$ is the estimator of τ_{jr}.

Final remarks

This paper has presented a method for two-point analysis that exploits quantitative information on dominant markers such as AFLPs. The simulations have shown that it will often be beneficial to exploit such information if it is available. In practice, the gain in efficiency has to be balanced against the cost of obtaining the quantitative data.

A referee suggested that there may be a break point at which dominant and co-dominant scoring might be equally efficient. The simulation results suggest that both methods become increasingly similar as the distance between means for the heterozygous and the dominant homozygous genotype diminishes. It does not seem, however, that dominant scoring may outperform co-dominant scoring, even if the overlap of the distributions underlying the quantitative variable is substantial. A probable reason is that dominant scoring always entails a loss of information, even if co-dominant scoring is less than perfect. Clearly, if co-dominant scoring is in error, then dominant scoring cannot be expected to fare any better. If, by contrast, the component distributions are very well separated, co-dominant scoring could even be done by eye with minimal classification error. In this situation, too, codominant scoring must be more informative. In summary, the more clearly separated are the means for the dominant homozygous and the heterozygous plants, the more pronounced is the gain in efficiency by co-dominant scoring. If separation is poor, then the gain is small or negligible. Simulations indicate that nothing is lost by trying to exploit quantitative information.

It is well known that the information content of dominant markers is especially low when dominant alleles are in repulsion linkage (Liu et al. 1998, p 193). Also, recombination-fraction estimates may be severely biased in this case, mainly due to the low frequency of the double-recessive class (Knapp et al. 1995). Therefore, it is to be expected that the gain by co-dominant scoring is even more pronounced for markers in repulusion than for markers in coupling. One reaction to the problem of the low-information content of dominant markers in repulsion has been to construct two separate maps. This approach exploits the fact that markers can be split into two groups of approximately equal size, with markers within each group linked in coupling. It is difficult to make the link between the two maps, however, since the repulsion linkage phase will be a barrier (Liu et al. 1998, p 193). For this reason, it is generally accepted that co-dominant markers are needed to anchor dominant markers. With the present method, it is possible in principle to construct a single map from all dominant markers, so the problem can potentially be resolved without resorting to anchoring. It would be interesting to compare the present method with map construction using dominant marker information in combination with anchoring markers. This question will be addressed in future work. Also note that extension of the present method to include co-dominant markers is straightforward, so anchoring can be integrated in the proposed method if desired.

References

Abramowitz M, Stegun A (1972) Handbook of mathematical functions. Dover Publications, New York

Basford KE, McLachlan GJ (1985) Estimation of allocation rate in a cluster analysis context. J Am Stat Assoc 80:286–293

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Jensen J, Jørgensen JH (1975) The barley chromosome-5 linkage map. Hereditas 80:5–16

Knapp SJ, Bridges WC Jr, Birkes D (1990) Mapping quantitative trait loci using molecular marker linkage maps. Theor Appl Genet 79:583–592

Knapp M, Wassmer G, Baur MP (1995) The relative efficiency of the Hardy-Weinberg equilibrium likelihood and the conditional on parental genotype-likelihood methods for candidate-gene association studies. Am J Hum Genet 57:1476–1485

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) MAPMAKER: an interactive computer package for constructing primary genetic maps of experimental and natural populations. Genomics 1:174–181

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Liu BH (1998) Statistical genomics. CRC Press, Boca Raton, Florida

Lynch M, Walsh B (1998) Genetical analysis of quantitative traits. Sinauer, Sunderland

McLachlan GJ, Basfrod KE (1988) Mixture models. Marcel Dekker, New York

McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley, New York

Ooijen JW van (1992) Accuracy of mapping quantitative trait loci in autogamous species. Theor Appl Genet 84:803–811

Ott J (1991) Analysis of human genetic linkage. John Hopkins University Press, Baltimore

Piepho HP, Koch G (2000) Codominant analysis of banding data from a dominant marker system by normal mixtures. Genetics 156:253–260

Stam P (1993) Construction of integrated genetic linage maps by means of a new computer package: JoinMap. Plant J 3:739–744

Weber WE, Wricke G (1994) Genetic markers in plant breeding. Parey Verlag, Berlin

Weir BS (1998) Genetic data analysis II. Sinauer, Sunderland

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1466