Letter to the Editor

A Comment on Codominant Scoring of AFLP Markers

Ritsert C. Jansen,* Henk Geerlings,† A. Jan Van Oeveren* and René C. Van Schaik†

*Centre for Biometry, Plant Research International B.V., NL-6700 AA, Wageningen, The Netherlands and
†Department of Bioinformatics, Keygene N.V., NL-6700 AE, Wageningen, The Netherlands

Manuscript received December 7, 2000
Accepted for publication March 14, 2001

PIEPHO and KOCH (2000) described a method for codominant analysis of banding data from a dominant marker system such as the amplified fragment length polymorphism (AFLP) technology (AFLP is a registered trademark of Keygene N.V.). Each AFLP marker detects the presence or absence of a specific fragment at a given genome location via gel or fluorescence methods. A strong point of the AFLP technology is that it allows for quantitative measurement of the degree of amplification of a fragment [termed band intensity here and optical density in PIEPHO and KOCH (2000)]. Individuals with one or two copies of the fragment show up, respectively, as “single-black” and “double-black” bands on a gel. In general a relation between copy number and quantitative measurement is expected and observed. Would this relation be linear? Would measurement errors be normal? Such questions need to be considered carefully to develop reliable models. Once we have a well-calibrated model, we can routinely use it for predicting copy numbers on the basis of quantitative band intensity measurements. This is very helpful for classifying individuals in one of the three genotypic classes in a segregating population obtained by crossing two diploid parents: homozygous AA (double band), heterozygous Aa (single band), or aa (no band). Unfortunately, the band intensity distributions of the three genotype classes may overlap, which hampers direct classification. Statistical mixture models can be used to overcome such problems, as nicely shown in the article by PIEPHO and KOCH (2000). Working on AFLP data, we too developed similar methods (GEERLINGS et al. 1999; our unpublished results, 1993). On the basis of what we learned so far from our data analyses, we think we can further the work presented by PIEPHO and KOCH (2000) by contributing in three essential ways.

(i) We provide evidence that band intensity values need to be square-root transformed prior to mixture modeling. This leads to distributions with constant variance within genotype classes AA, Aa, and aa on the transformed scale.

(ii) We present a simple approach controlling misclassification errors. This is of importance for many marker-based applications in which erroneous marker scores introduce false recombinant genotypes.

(iii) We propose a histogram visualization. The histogram can be used to form an opinion on the goodness-of-fit of the model to the data and to quickly eliminate problematic markers.

Figure 1a shows a histogram of band intensities of 87 tomato plants from a segregating F2 progeny. We clearly see a trimodal distribution, with a mode for each of the three genotype classes—aa, Aa, and AA. The mean values are close to the ratio 0:1:2; that is, band intensity is linearly related to copy number. We also see that the variances around the modes increase linearly with the means: narrow for aa, intermediate for Aa, and widest for AA. In such situations square root is the transformation of choice.

Next we fit normal mixture models to square-root-transformed data. We refer to Jansen (1993) and Jansen (2001) for a detailed description of procedures for fitting (generalized linear) mixture models and plotting back-transformed mixture distributions on top of the histogram. Figure 1a shows clearly that the mixture models fit well to the data, in terms of mean as well as in terms of variance. We analyzed many other AFLP markers and obtained similar good fits (not shown). Our histogram visualization is more informative than the visualization in Figure 3 of PIEPHO and KOCH (2000). Their Figure 3 also shows that they worked with higher levels of background noise, which limited the power for discrimination between various transformation options in their study.

Figure 1a also shows overlap between the distribution of Aa and AA. We strongly recommend not classifying individuals with band intensities in the overlapping regions. Instead, we propose to add two more classes, not AA and not aa, for the first and second overlap regions,
Moving from the left to the right, we classify individuals as aa, not AA, Aa, not aa, and AA. The four vertical lines in Figure 1a define the five genotype regions. The lines are set in such a way that the posterior probability of really having genotype aa in the “aa region” is at least 98% (similarly for Aa and AA, respectively). The two other regions are “regions of doubt.”

Piepho and Koch (2000) analyzed data for 46 individuals. Our Figure 1b shows a histogram of 45 F2 tomato plants, obtained from one-half of a gel, and clearly the model fits well to the data. In this smaller example segregation is clearly distorted. In such cases estimating the genotype mixing proportions is preferable to using the theoretical proportions 1:2:1. The genotype mixing proportions are estimated as 0.47, 0.40, and 0.13 in Figure 1b.

This note does not intend to disqualify the method described by Piepho and Koch (2000). Instead our note adds further information highly essential for many practical applications, but still missing in their article. Our results are particularly important for the codominant use of AFLP markers in high-density linkage analysis, mapping of quantitative trait loci, and marker-assisted breeding.

LITERATURE CITED


Communicating editor: R. G. Shaw