BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 16 2010, pages 1983–1989
doi:10.1093/bioinformatics/btq316

Gene expression

Advance Access publication June 23, 2010

Single feature polymorphism detection using recombinant inbred line microarray expression data

Xinping Cui[1,2,*], Na You[1], Thomas Girke[2,3], Richard Michelmore[4] and Allen Van Deynze[5]

[1]Department of Statistics, [2]Institute for Integrative Genome Biology and Center for Plant Cell Biology, [3]Department of Botany and Plant Sciences, University of California, Riverside, [4]Genome Center and Department of Plant Sciences and [5]Seed Biotechnology Center, University of California, Davis, CA, USA

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The Affymetrix GeneChip® microarray is currently providing a high-density and economical platform for discovery of genetic polymorphisms. Microarray data for single feature polymorphism (SFP) detection in recombinant inbred lines (RILs) can capitalize on the high level of replication available for each locus in the RIL population. It was suggested that the binding affinities from all of the RILs would form a multimodal distribution for a SFP. This motivated us to estimate the binding affinities from the robust multi-array analysis (RMA) method and formulate the SFP detection problem as a hypothesis testing problem, i.e. testing whether the underlying distribution of the estimated binding affinity (EBA) values of a probe is unimodal or multimodal.

Results: We developed a bootstrap-based hypothesis testing procedure using the 'dip' statistic. Our simulation studies show that the proposed procedure can reach satisfactory detection power with false discovery rate controlled at a desired level and is robust to the unimodal distribution assumption, which facilitates wide application of the proposed procedure. Our analysis of the real data identified more than four times the SFPs compared to the previous studies, covering 96% of their findings. The constructed genetic map using the SFP markers predicted from our procedure shows over 99% concordance of the genetic orders of these markers with their known physical locations on the genome sequence.

Availability: The R package 'dipSFP' can be downloaded from http://sites.google.com/a/bioinformatics.ucr.edu/xinping-cui/home/software

Contact: xinping.cui@ucr.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 3, 2009; revised on May 5, 2010; accepted on June 10, 2010

1 INTRODUCTION

In the past decade, Affymetrix GeneChip® microarrays have provided an important and versatile tool for many genome studies. This technology uses hybridization probes which are comprised of hundreds of thousands of 25-mer oligonucelotides chemically synthesized on a solid substrate. Typically, these probes (or features) are grouped into different probe sets for different target genes (see Affymetrix 2001 for more details). Since its introduction by Lockhart *et al.* (1996), Affymetrix GeneChip® microarrays have been used extensively for genome-wide gene expression analysis.

Recently, Affymetrix GeneChip® microarrays have also been used for identifying genetic polymorphisms on a variety of genomes in a highly parallel manner (Banks *et al.*, 2009; Borevitz *et al.*, 2003; Cui *et al.*, 2005; Das *et al.*, 2008; Kumer *et al.*, 2007; Ronald *et al.*, 2005; Rostocks *et al.*, 2005; Wang *et al.*, 2009; West *et al.*, 2006; Winzeler *et al.*, 1998). Such genetic polymorphisms, including single nucleotide polymorphisms (SNPs) and small insertion/deletions, are detected by single probes on the array and, therefore, are called single feature polymorphisms (SFPs). The idea is that the sequence mismatch between a 25-mer perfect match (PM) probe on the array and its target sequence will significantly alter hybridization and attenuate that probe's signal. Therefore, the presence of sequence polymorphisms in some PM probes will lead to detectable differences in those probes' hybridization signals between the two different accessions, one of which has similar sequences as the probe sequences on the array. Since hundreds of thousands of probes can be analyzed at the same time, the Affymetrix GeneChip® microarrays are currently providing one of the most highly parallelized and economical platforms for genetic polymorphism discovery, facilitating the increasing need of high-resolution genetic and physical maps for genetic studies (Borevitz *et al.*, 2007; Jiang *et al.*, 2006; Kim *et al.*, 2006; Werner *et al.*, 2005). For relatively small genomes, such as yeast, arabidopsis and rice, microarray hybridization of genomic DNA samples has been a successful route to detect SFPs (Borevitz *et al.*, 2003; Kumar *et al.*, 2007; Ronald *et al.*, 2005; Winzeler *et al.*, 1998). For more complex genomes, such as barley and cowpea, cRNA samples have been successfully used as surrogates for DNA samples to detect SFPs (Banks *et al.*, 2009; Cui *et al.*, 2005; Das *et al.*, 2008; Rostocks *et al.*, 2005).

In most of the microarray-based SFP detection studies, usually only two or three biologically replicated DNA or RNA samples are extracted from each accession under comparison. With so few replicates, SFP detection suffers from either low detection power or high false positive rate. However, increasing biological replicates simply for the purpose of SFP detection is generally cost prohibitive. Since RIL expression microarrays are often available for expression quantitative trait locus analysis, West *et al.* (2006) proposed to use such data, without extra cost, for simultaneous SFP detection and RIL genotyping. They developed a metric for the binding affinity called SFPdev, and assumed that the SFPdev values of a SFP from

*To whom correspondence should be addressed.

all of the RILs would form a bimodal distribution, where each peak represents a parental genotype. Then for each probe, they sorted the SFPdev values in an increasing order and performed an empirical brute force search for a gap value greater than two in the ratios of every two consecutive SFPdev values. If such a gap was found for a probe and all SFPdev values of one parent fall on one side of the gap and all SFPdev values of the other parent fall on the other side of the gap, then this probe is called a SFP. Obviously, a simple ratio of two consecutive SFPdev values cannot fully characterize the distribution of SFPdev values from all of the RILs, and the searching method also lacks the statistical justification of false positive control. Moreover, it has been reported in the literature (Luo *et al*., 2007) that most of SFPs predicted from the SFPdev-based analysis are gene expression level polymorphisms rather than the genuine sequence polymorphisms, suggesting that the SFPdev is not independent of the gene expression level. In this article, we formulate the SFP detection problem as a hypothesis testing of unimodality versus multimodality, and propose a bootstrap testing procedure to predict SFPs based on the estimated binding affinity (EBA) values from the robust multi-array analysis (RMA) method.

2 METHODS

2.1 Probe binding affinity index

The PM intensity value obtained from the microarray is a compound of multiple effects. Irizarry *et al*. (2003) proposed to decompose the PM values of the probes within one probe set as following:

$$\mathrm{PM}_{kj} = \phi_k \theta_j \epsilon_{kj} + B_{kj}, \tag{1}$$

where ϕ_k is the binding affinity between the target sequence and the k-th probe, θ_j the gene expression level of the probe set on the j-th array, ϵ_{kj} the random error and B_{kj} represents the background noise. SFP detection emphasizes on the sequence polymorphism prediction, which is reflected by the different binding affinities between target sequences bearing and not bearing sequence polymorphism(s) when hybridized on the same probes on the arrays. For the robustness of SFP detection to the major variation in the gene expression data, it is essential to have an index that can measure the binding affinity and is independent of the gene expression level.

West *et al*. (2006) developed a metric, called SFPdev, to estimate the binding affinity from the PM values. For probe i, it is calculated by '|PM intensity for probe i-average of PM intensities for the remaining 10 probes in the same probe set|/PM intensity for probe i'. Luo *et al*. (2007) proposed three indexes incorporating the mismatch (MM) probe values, which are $I_1 = \log(\mathrm{PM}/\mathrm{MM})$, $I_2 = (\mathrm{PM}-\mathrm{MM})/\mathrm{PM}$ and $I_3 = (\mathrm{PM}-\mathrm{MM})/\mathrm{MM}$. In this article, we use the EBA from model (1) as the measurement of the true binding affinity ϕ_k. The EBA values of the probes in a given probe set can be calculated by the Tukey's median polish method from the background corrected (Chen *et al*., 2009), quantile normalized (Irizarry *et al*., 2003) and log-transformed PM values on the replicated arrays. The obtained EBA values are then summarized in a matrix $\{X_{ij}\}_{m\times n}$, where each row corresponds to one probe and each column corresponds to one RIL. Further SFP detection procedures are performed on the dataset $\{X_{ij}\}_{m\times n}$.

2.2 Hypothesis testing

According to West *et al*. (2006), in the homozygous RIL population, the distribution of the binding affinities is bimodal for an SFP and unimodal for a non-SFP. More generally, it could also be trimodal for a SFP locus if the RIL population is heterozygous and rarely multimodal if a mutation had occurred during generation of the RILs. In other words, X_{ij} in each row could follow either a multimodal distribution or a unimodal distribution depending on whether the row corresponds to a SFP or not. Therefore, the problem of determining whether a probe, e.g. probe i, is a SFP or not can be formulated as testing the following hypothesis:

$$H_{i0}: f_i(x) \text{ is unimodal versus } H_{i1}: f_i(x) \text{ is multimodal}, \tag{2}$$

where $f_i(x)$ is the underlying density function of the i-th row in the matrix $\{X_{ij}\}_{m\times n}$. For testing hypothesis (2), several statistics have been proposed in the literature, among which the most commonly used and intensively studied are the Silverman critical bandwidth statistic (Silverman, 1981), the dip statistic (Hartigan and Hartigan, 1985) and the excess mass statistic (Müller and Sawitzki, 1991). Although the Silverman test is heavily cited and has been called 'ingenious' (Fisher and Marron, 2001), previous studies suggested three major drawbacks of the critical bandwidth statistic: (i) it has poor distributional properties; (ii) it is very sensitive to outliers; and (iii) it makes no distinction between tiny, 'negligible' modes and large, 'important' modes. For 1D data, the dip statistic and the excess mass statistic are equivalent (Cheng and Hall, 1998). Therefore, we choose the dip statistic for the multimodality test of the distribution of X_{ij} obtained from all of the RIL microarrays.

Given a random sample $X_1, X_2, \ldots, X_n$ from an unknown probability density function $f(x)$, the dip statistic is defined as the maximum difference between the empirical distribution function and the unimodal distribution function which minimizes that difference. It mathematically can be written as (Hartigan and Hartigan, 1985),

$$D = \inf_{G\in\mathcal{U}} \sup_x |\hat{F}_n(x) - G(x)|, \tag{3}$$

where $\mathcal{U}$ is the class of cumulative distribution functions with unimodal densities and $\hat{F}_n(x)$ the empirical cumulative distribution function. When $f(x)$ is uniform on $(0,1)$, the dip statistic D is obtained given $G(x)$ being uniform on $(0,1)$ and the asymptotic distribution of dip statistic D can be derived. The percentile points of D in the finite sample size can also be tabulated via simulation (Hartigan and Hartigan, 1985). However, when $f(x)$ is unimodal but not uniform, the asymptotic distribution of D is not known.

To obtain the null distribution of the dip statistic D under the unimodal distribution assumption, we employed the smooth bootstrap algorithm (Efron and Tibshirani, 1993) and the modified critical bandwidth (Fisher and Marron, 2001). The smooth bootstrap is equivalent to re-sampling from a kernel density estimate. For the random sample $X_1, X_2, \ldots, X_n$ from the probability density function $f(x)$, the kernel density estimate $\widehat{f}_h(x) = \sum_{j=1}^n K_h(x - X_j)/n$, where $K_h(.) = K(\cdot/h)/h$ is a re-scaling of the kernel function K, which is chosen to be the standard Gaussian probability density in this article, and h is the bandwidth. Given the tuning parameters λ_0 and m_0, the modified critical bandwidth h_k in Fisher and Marron (2001) is the minimal bandwidth at which $\widehat{f}_{h_k}(x) - \lambda_0$ has k 'major' modes whose masses are larger than m_0. Specifically, the density $\widehat{f}_{h_1}$ is the kernel density estimate with one 'major' mode that is closest to the data and h_1 is called the critical bandwidth for one mode. As suggested by Fisher and Marron (2001), the modified critical bandwidth can address the drawbacks (ii) and (iii) of the critical bandwidth that we mentioned earlier. It is reasonable to use $\widehat{f}_{h_1}$ to estimate $f(x)$ under the assumption that $f(x)$ is unimodal. However, $\widehat{f}_{h_1}$ has to be rescaled to have variance equal to the sample variance. Denote the rescaled $\widehat{f}_{h_1}$ by $\widehat{g}_{h_1}$. Because of the convenient form of the Gaussian kernel estimate, drawing samples from $\widehat{g}_{h_1}$ can be easily done. Denoting the observed data by $x_1, x_2, \ldots, x_n$, we can sample $y_1^*, y_2^*, \ldots, y_n^*$ with replacement from $x_1, x_2, \ldots, x_n$ and set

$$x_j^* = \bar{y}^* + (1 + h_1^2/\widehat{\sigma}^2)^{-1/2}(y_j^* - \bar{y}^* + h_1\epsilon_j),\ j = 1, 2, \ldots, n, \tag{4}$$

where $\bar{y}^*$ is the mean of $y_1^*, y_2^*, \ldots, y_n^*$, $\widehat{\sigma}^2$ the plug-in estimate of variance of the data and ϵ_j the standard normal random variables. Putting all of these together, our bootstrap hypothesis test for SFP detection using dip statistic D is summarized as following:

1. Calculate the dip statistic D for each probe i using formula (3), and denote it by D_i.
2. Calculate the critical bandwidth h_1 for each probe i, denoted by h_{1i}, $i = 1, 2, \ldots, m$.

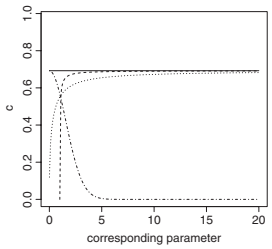

Fig. 1. Solid: $N(\mu, \sigma^2)$; dashed: gamma(k, θ); dotted: $T(\nu)$; dotdash: log-normal (μ, σ^2). Corresponding parameter stands for μ or σ for $N(\mu, \sigma^2)$, k for gamma (k, θ), ν for $T(\nu)$ and σ for log-normal (μ, σ^2).

3. For each probe i, draw B bootstrap samples of size n from $\widehat{g}_{h_{1i}}$ using (4) and calculate dip statistics $D^*_{i1}, D^*_{i2}, \ldots, D^*_{iB}$, $i = 1, 2, \ldots, m$.

4. Approximate P-value of probe i by

$$\widehat{p}_i = \frac{\#\{D^*_{lk} \geq D_i, l = 1, 2, \ldots, m; k = 1, 2, \ldots, B\}}{mB}$$

5. Adjust P-values $\widehat{p}_i$, $i = 1, 2, \ldots, m$, using the false discovery rate criterion (Benjamini and Hochberg, 1995). Given a pre-determined α, probes with adjusted P-values smaller than α are selected as SFPs.

Cheng and Hall (1998) studied the asymptotic property of the dip statistic and concluded that, under condition (2.2) outlined in their paper, the asymptotic null distribution of the dip statistic is independent of the unknown $f(x)$ except for a factor $c = \left\{ f(x_0)^3 / |f''(x_0)| \right\}^{1/5}$, where x_0 is the unique mode of $f(x)$. The specific formulas of factor c for normal, gamma, t and log-normal distributions are listed in Supplementary Table S1. Figure 1 provides the graphic illustration of how the value of c changes as the dependent parameters change. As can be seen, if $f(x)$ is a normal density, then $c = (2\pi)^{-1/5}$ is a constant and reflected by a flat line in Figure 1. This means for any two normal densities with different means and variances, their dip statistics converge to the same asymptotic distribution as the sample size n goes to infinity. Although for gamma and t densities, the corresponding factor c depends on one of their density parameters, such dependency quickly diminishes as the value of dependent parameter increases. Moreover, with moderately large values of dependent parameters (shape parameter for gamma density and degrees of freedom for t density), the values of factor c and, therefore, the asymptotic null distributions of the dip statistics are very close to that of normal density. For log-normal density, when σ^2 is small, the values of factor c and the asymptotic null distributions of the dip statistics are still close to that of the normal density. However, as σ increases, the values of factor c and, therefore, the asymptotic null distributions of the dip statistics are farther away from that of the normal density, although the dependency between σ^2 and c quickly disappears. For this case, we suggest to perform a log-transformation on the observed data before bootstrap resampling in Step 3 and then take the exponential function on resampled data to calculate the bootstrap version dip statistics. Supplementary Figure S1 shows the distributions of the dip statistics when $f(x)$ is normal, gamma, t or log-normal density with varying parameters. The densities are plotted based on 10 000 replicated dip statistic values, each of which was calculated through 100 observations sampled from $f(x)$. As shown in Supplementary Figure S1, the distributions of the dip statistics weakly depend on the corresponding density parameters that are moderately large for gamma and t densities and small for log-normal density. Throughout the article, for all of the probes we assume the distributions of the EBA values from all of the RILs follow the *same family of unimodal distributions under the null hypotheses.* The above results provide us the ground to assume a common null distribution for dip statistics $D_1, D_2, \ldots, D_m$. Therefore, we propose to pool all the bootstrap version dip statistics from m probes to estimate the null distribution of the dip statistic in Step 4 of our testing procedure.

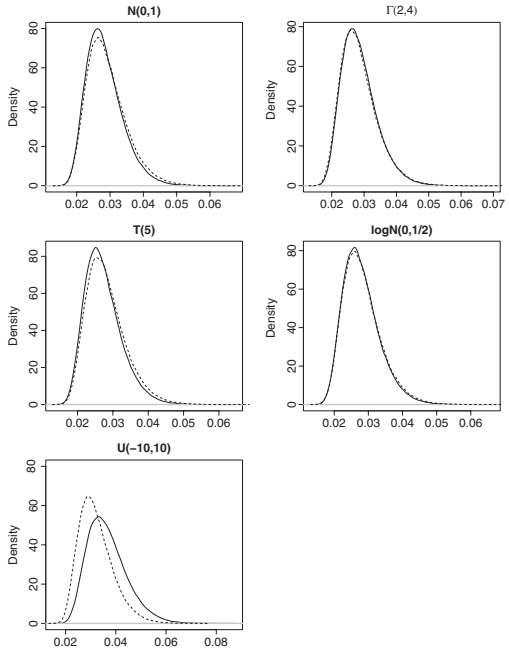

Fig. 2. True (solid) and estimated (dashed) null distribution of the dip statistic when the null underlying density $f_i(x)$ is normal, gamma, t, log-normal or uniform density.

3 RESULTS

3.1 Simulation Study

A series of simulations were performed to assess the performance of our proposed algorithm. In each simulation study, the sample size n is always set to be 100 and the total number of probes $m = 10\,000$, among which there are 500 SFPs generated from the mixtures of two distributions from the same family with the mixing probability p. Tuning parameter λ_0 is set to be 0, $m_0 = 0.003$ as suggested by Fisher and Marron (2001). The number of bootstrap replications B is set to be 20 in all simulation studies. Five commonly used density families, including normal, gamma, t, log-normal and uniform, are examined in the simulation studies (see Supplementary Table S2 for the detailed density functions of these density families used for data generation).

The first simulation is to evaluate the performance of the bootstrap procedure in estimating the null distribution of the dip statistic. We considered five scenarios corresponding to the five density families mentioned above. The parameter settings in five scenarios are given in Supplementary Table S3. In each scenario, $P = 0.2$, 0.3, 0.4, 0.5 and 0.6 were considered. Note that for all non-SFPs, data were generated from the same unimodal distribution. For SFPs, data were generated from the mixtures of distributions from the same family with the density parameters varying randomly in order to reflect different levels of separation in two modes of the underlying distributions among SFPs. The true and estimated null distributions of the dip statistics are presented in Figure 2. Since the estimated null distributions with different P settings are very similar, only the estimated null distribution with $P = 0.5$ are plotted in Figure 2. It can be seen that the null distributions of the dip statistics are well estimated by our proposed bootstrap procedure for normal density, gamma and t densities with moderately large dependent parameters

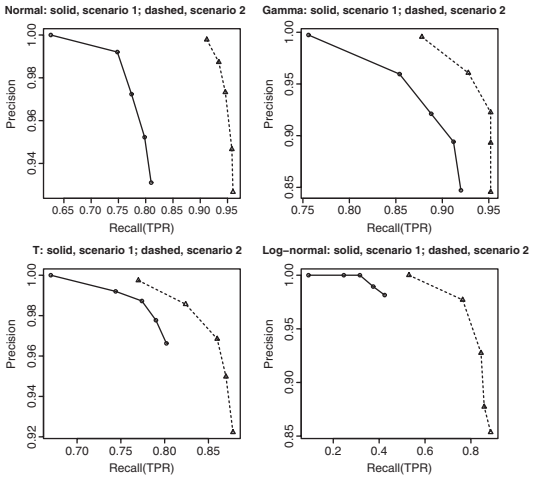

Fig. 3. PR curves calculated from our proposed algorithm when $\alpha = 0.01, 0.05, 0.1, 0.15$ and 0.2 in the case that the null underlying density $f_i(x)$ is normal, gamma, t, log-normal or uniform density.

and log-normal density with small σ. The estimated null distribution of the dip statistic for t density has a positive bias from the truth when the degrees of freedom are small (data not shown), and the estimation bias becomes smaller and smaller as the degrees of freedom increase. We observed a similar behavior for the log-normal density, except the bias reduces as the parameter σ decreases and it is more sensitive to the change of dependent parameter values. For the uniform case, the bias is obviously negative and we suggest to use the percentile table in Hartigan and Hartigan (1985) to determine the significances of dip statistic values and, therefore, exclude it in the following simulation.

The second simulation is to evaluate the false discovery rate (FDR) and detection power of the proposed algorithm. To reflect the real situation that for different non-SFPs, the underlying density functions $f_i(x)$ could be in the same family but with different parameters, the data in this simulation are generated under the parameter settings in Supplementary Table S4. The precision–recall (PR) curves of eight scenarios described in Supplementary Table S4 are plotted in Figure 3 when $\alpha = 0.01, 0.05, 0.1, 0.15$ and 0.2, where the precision (=1−FDR) is the proportion of true positives among total number of rejections, and the recall (=1−Type II error rate) is the ratio of the number of true positives to the total number of true positives and false negatives. In the context of SFP discovery, 'Precision' measures the fraction of probes classified as SFPs that are truly SFPs, while 'Recall' measures the fraction of total true SFPs that can be discovered. As shown in Figure 3, the FDR of our procedure is controlled under the desired level. With consideration of the conservativeness of the FDR control procedure of Benjamini and Hochberg (1995), they are still in reasonable range except for Scenario 1 of t and log-normal densities, where the actual FDRs are much smaller than α. In these two cases, the degrees of freedom of the t density are set to be quite small and the σ parameters of log-normal density are set to be large, resulting in positive estimation bias of the null distribution of the dip statistic and, therefore, larger P-values. The detection power of our procedure depends on the separation of two modes in the mixture distributions of SFPs. Within each density family, the two modes of SFPs in the second scenario are better separated than that in the first scenario, resulting in higher detection power of Scenario 2 (dashed) than Scenario 1 (solid) at the same level of FDR control.

To evaluate the performance of our procedure for SFP detection in the population with distorted segregation ratios, four more scenarios, Scenario 3–6, are considered for each density family. The parameter settings are listed in Supplementary Table S5. As shown by the PR curves in Supplementary Figure S2, for SFPs with small modes in the bimodal distributions (Scenario 3 of the normal, gamma and t densities), our detection power is very low. In those cases, it is hard to say whether the small mode appearing in the data is randomly generated from a unimodal distribution or caused by a bimodal distribution. Although this feature reduces the detection power, it justifies the robustness of our procedure to the outliers which is often an issue in the real data analysis. In Scenario 6 of the log-normal density where much flatter modes are generated for some SFPs compared to the others in the bimodal distributions, our procedure treats the flat modes like heavy tails and classifies them as non-SFPs, which also results in a low detection power.

3.2 Real data analysis

The gene expression microarray dataset consisting of two biological replicates of 148 *Bay-0*×*Sha* F9 RILs of *Arabidopsis thaliana* and 8 parental accessions (4 from *Sha* and 4 from *Bay-0*) was obtained from West *et al*. (2006). We only used the PM values. After background correction and normalization, the replicated data are stored in 150 matrices, of which, 148 are of $m \times 2$ corresponding to 148 RILs with two replicates each and 2 are of $m \times 4$ corresponding to two parental controls with four replicates each, where $m = 249\,975$ indicates the number of probes on the arrays. Note that only the probe sets containing 11 probes were considered in the analysis. The EBA values are then obtained and summarized in the matrix $(X_{ij})_{m\times 148}$ as described in Section 2.1, where each row represents a probe and each column corresponds to a RIL. Given $\alpha = 0.01$, 0.001, 1e-5 or 1e-10, the number of SFPs predicted from our procedure is 7851, 6534, 5492 or 5225, respectively. The adjusted P values of the majority of predicted SFPs are smaller than 1e-10, which actually equals to zero. Therefore, we set the nominal FDR control α = 1e-10, resulting in 5225 predicted SFPs.

To assess our SFP detection results, we retrieved the probe sequences from the Bioconductor package `ath1121501probe` and mapped them onto chromosome sequences. The results were compared with the known SNPs of the genotypes *Sha* and *Bay-0* (from `ftp://ftp.arabidopsis.org/Sequences/JGI_Resequencing_Data`) by searching those probes containing SNPs. In total, 231 880 probes mapped to chromosome locations, covering 255 065 positions. To obtain unambiguous results, only the probes with unique mappings were considered in this study. Moreover, we only considered SNPs that occurred either in *Sha* or *Bay-0* relative to *Col-0*, since it is still not clear what the hybridization profile of heterozygotes or mixed genotypes (SNP in both) is on microarrays. Among 220 184 uniquely matched probes, 15 856 probes cover SNPs. For 5225 SFPs predicted from our procedure, 4745 (distributed in 2730 probe sets) are uniquely located on chromosomes. Their overlap with the probes covering SNPs is 2989, resulting in 37% actual FDR. Xie *et al*. (2009) showed that the false positive SFPs can result from SNPs adjacent to probe binding regions. Of 1756 false positives, there are 499 probes whose flanking

Table 1. The numbers of predicted SFPs, detection powers and FDRs calculated from different indexes

	Number of SFPs	Detection power	FDR[1]	FDR[2]
EBA	4745	0.19	0.37	0.26
SFPdev	3205	0.13	0.34	0.24
I_1	4883	0.19	0.38	0.31
I_2	3885	0.15	0.39	0.31
I_3	4159	0.16	0.37	0.29

FDR[1], FDR without considering the flanking SNPs; FDR[2], FDR with consideration of SNPs within flanking 20 bp.

20 bp sequences cover SNPs, reducing the previous seemingly large FDR to 26%. A bit smaller FDR (24.4%) was reported in Xie *et al*. (2009). However, it was calculated based on a subset of predicted SFPs and the FDR for the whole set of predicted SFPs is not known. The expression level polymorphism also might cause the false discovery. Luo *et al*. (2007) concluded that SFPs predicted from the Affymetrix microarrays mainly (∼64%) represents the polymorphisms in *cis*- or *trans*-acting regulators rather than the sequence polymorphisms in the barley genome studies. If there are multiple probes identified as SFPs in one probe set, they are more likely due to the expression level polymorphisms rather than the sequence polymorphisms. The remaining 1257 false positives distribute in 805 probe sets, among which 16 probe sets contained six or more SFPs, covering 108 (8.6%) false positive probes.

SFP detection is built on the idea that the target sequences bearing polymorphism(s) will bind poorly to the corresponding probes, resulting in significantly lower hybridization signal intensities compared to that of the target sequences that are highly similar to the probe sequences. Sometimes, however, even if the target sequences are highly similar to the probe sequences, we could still observe poor bindings, making the SFP detection infeasible for those probes. We used the 90 percentile of the PM values of one probe as the measurement of the hybridization signal intensities of this probe, and compared that of the 2989 probes that bear SNPs and were detected by our procedure with that of the 12 589 probes that bear SNPs but were not detected by our procedure. The 20 percentile of the former (153) is even larger than the 80 percentile of the later (130), suggesting the majority of probes bearing SNPs were missed by our procedure due to the poor bindings at these probes. The detection power is also related to the SNP locations in the probes. For the probes bearing multiple SNPs, we only count the SNP closest to the middle (the 13th base) of 25-mer sequence. Using our procedure, the detection rate of SFPs whose SNPs locate in the middle is 27%, while the detection rate of SFPs bearing SNPs in the edge (the 1st and 25th bases) is 2.5%.

To compare the performances of different binding affinity indexes, we re-ran the proposed procedure with all of the indexes mentioned in Section 2.1. The numbers of predicted SFPs, the detection power and the FDRs with and without consideration of flanking SNPs under the same nominal FDR control (α = 1e-10) are listed in Table 1. As shown in Table 1, the EBA reaches the highest detection power with the actual FDR controlled at the comparable level with the SFPdev, whose actual FDR is the lowest. West *et al*. (2006) identified 1257 SFPs, of which 1148 probes (distributed in 895 probe sets) are uniquely matched on the chromosomes. It is worth mentioning that

Table 2. The numbers of SFPs and genetic bins in each linkage group

Linkage group	1	2	3	4	5
Number of SFPs	704	395	529	408	661
Number of bins	226	153	169	145	251
Map length (cM)	187.82	126.83	134.50	112.79	205.37
Average distance[a] (cM)	0.83	0.83	0.80	0.78	0.82

[a]Average genetic distance between bins.

96% (1103/1148) of SFPs or 99.6% (891/895) of probe sets identified by West *et al*. (2006) are also detected by our procedure.

The genetic map was generated using the identified SFPs. First, for each SFP, the EBA values from all of the RILs are clustered into two groups by the fuzzy clustering method (Kaufman and Rousseeuw, 1990) and the EBA values of parental controls used to determine which parental genotype each group belongs to. If the EBA value of *Sha* is larger than that of *Bay-0*, then the RILs clustered in the group with larger EBA values are all genotyped as *Sha* and the RILs in the other group are all genotyped as *Bay-0*. Otherwise, the genotypes are assigned reversely. Note that a genotype is assigned to a RIL only if its probability of being in one of the two clusters is greater than 0.7 which is calculated from the fuzzy clustering procedure, otherwise, the genotype is assigned as missing. Second, for the probe sets containing multiple SFPs, one SFP with the minimum missing genotype scores is selected per probe set. Third, the redundant SFPs are filtered using the MadMapper software (http://cgpdb.ucdavis.edu/XLinkage/MadMapper/) and removed from further analysis. SFPs with 25% or more missing genotype scores are also removed. The MSTmap software clusters the resulted 2697 SFPs into five linkage groups, and within each group, the SFPs are grouped into genetic bins and sorted according to their genetic orders. The numbers of SFPs and genetic bins in each linkage group are presented in Table 2, together with the length of the constructed genetic map and the average genetic distance between bins in each linkage group. The SFP markers we detected offer more complete genome coverage than that of West *et al*. (2006). In the previous sequence analysis, all the uniquely matched probes were located on the chromosomes. We plotted the physical locations versus genetic positions of predicted SFPs from the MSTmap analysis (Fig. 4). As shown in Figure 4, there are 3, 1, 8, 3 and 9 SFPs in five linkage groups whose genetic positions are not consistent with their physical locations (Supplementary Table S6). In other words, we observed over 99% concordance of the genetic orders of the predicted SFPs with their known physical locations on the genome sequence. There are four inconsistent SFP markers found in West *et al*. (2006), which are all included in the above 24 inconsistent SFPs. Among the remaining 20 inconsistent SFPs, 6 belong to the probe sets that correspond to multi-gene families.

4 DISCUSSION

Our simulation studies show that our procedure can reach a satisfactory detection power with FDR controlled at the desired level when there are sufficiently strong signal intensities from the microarray hybridization. They also show that our procedure is robust to the unimodal distribution assumption, which allows the broader application of our procedure. Our analysis of the RIL data

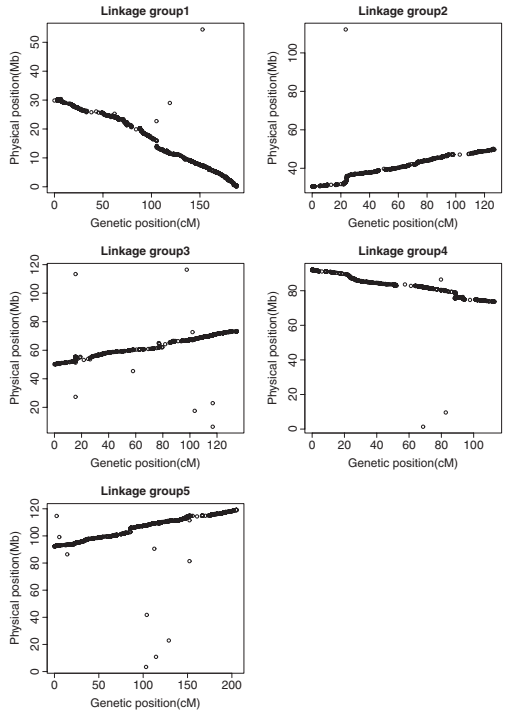

Fig. 4. Plots of genetic positions versus physical locations of SFPs in five linkage groups.

of West *et al*. (2006) detected more than four times as many SFPs, covering 96% of that of West *et al*. (2006). The constructed genetic map using the predicted SFPs showed over 99% concordance of the genetic positions with their known physical locations on genome sequence.

To account for the binding affinity from the PM intensity value, we proposed the EBA, which is estimated by the Tukey's median polish method from the RMA model (1). However, in the real data analysis, we obtained the EBA by dividing the PM value by the estimated expression level, since there were only two replicates for each RIL and the estimate for the binding affinity might not be reliable. The performance of the EBA was investigated in the real data analysis. Compared to the low detection power of SFPdev and high FDRs of indexes proposed in Luo *et al*. (2007), the EBA showed the advantage of high detection power and competitive FDR control.

Because of the conservativeness of the FDR control procedure of Benjamini and Hochberg (1995), the actual FDR from our procedure is usually smaller than the nominal FDR. However, in the real data analysis, a seemingly larger FDR was obtained. This phenomenon has also been reported in the literature. In additional to the binding affinity and the gene expression level, many factors could affect the signals on the arrays and result in the SFP like probes that cannot be validated by the sequence polymorphisms (Xie *et al*., 2009). In our simulations where $f_i(x)$ is multimodal if and only if the probe i is a SFP, our algorithm works well and the FDR is controlled at the nominal level.

Note that the detection power highly depends on the separation of two modes in the bimodal distribution. The low detection power in our real data analysis are mainly due to the poor bindings of the target sequences that perfectly match to the probe sequences on the array, resulting in the modes not separatable. The on-going efforts on SFP technology (Gupta *et al*., 2008; Gore *et al*., 2007), including complexity reduction and gene enrichment of the target DNA, can potentially improve the detection power of SFP detection.

This article is motivated by the SFP detection predominantly for homozygous RIL populations, where there are only two genotypes at most loci. Rejecting the null hypothesis directly implies the sequence difference between those two genotypes. In other settings, there might be more than two possible genotypes, resulting in bimodal, trimodal or multimodal distribution of the EBA values for a SFP. The number of modes is of great interest for genotype identification, which can then be formulated as the following hypothesis testing:

$$H_0 : j \leq k \text{ versus } H_1 : j > k,$$

where j is the number of modes of $f_i(x)$ and $k = 1, 2, 3$ or more. Fisher and Marron (2001) proposed a statistic T_k for this hypothesis testing and used the smooth bootstrap method to generate the null distribution of T_k to measure the significance of the statistic. To determine the number of modes for SFPs, we can apply this procedure for $k = 1, 2, 3, \ldots$ until the null hypothesis is accepted. The k at which the null hypothesis is accepted is an estimate of the number of modes.

Funding: The National Science Foundation DBI 0646024 (to X.C. and N.Y.).

Conflict of Interest: none declared.

REFERENCES

Banks,T.W. *et al*. (2009) Single-feature polymorphism mapping in bread wheat. *Plant Genome*, **2**, 167–178.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Borevitz,J.O. *et al*. (2007) Genome-wide patterns of single-feature polymorphism in arabidopsis thaliana. *Proc. Natl Acad. Sci. USA*, **104**, 12057–12062.

Borevitz,J.O. *et al*. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.*, **13**, 513–523.

Chen,Z., *et al*. (2009) A distribution-free convolution model for background correction of oligonucleotide microarray data. *BMC Genomics*, **10** (Suppl. 1), S19.

Cheng,M.-Y. and Hall,P. (1998) On mode testing and empirical approximations to distributions. *Stat. Probab. Lett.*, **39**, 245–254.

Cui,X. *et al*. (2005) Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit. *Bioinformatics*, **21**, 3852–3858.

Das,S. *et al*. (2008) Detection and validation of single feature polymorphisms in cowpea (vigna unguiculata l. walp) using a soybean genome array. *BMC Genomics*, **9**, 107.

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

Fisher,N.I. and Marron,J.S. (2001) Mode testing via the excess mass estimate. *Biometrika*, **88**, 499–517.

Gore,M. *et al*. (2007) Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. *Crop Sci.*, **47**, S135–S148.

Gupta,P.K. *et al*. (2008) Array-based high-throughput dna markers for crop improvement. *Heredity*, **101**, 5–18.

Hartigan,J.A. and Hartigan,P.M. (1985) The dip test of unimodality. *Ann. Stat.*, **13**, 70–84.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Jiang,R. *et al*. (2006) Inferring population parameters from single-feature polymorphism data. *Genetics*, **173**, 2257–2267.

Kaufman,L. and Rousseeuw,P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. 9th edn. Wiley-Interscience, New York.

Kim,S. *et al*. (2006) Association mapping with single-feature polymorphisms. *Genetics*, **173**, 1125–1133.

Kumar,R. *et al*. (2007) Single feature polymorphism discovery in rice. *PLoS ONE*, **2**, e284.

Lockhart,D.J. *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Luo,Z.W. *et al*. (2007) SFP genotyping from Affymetrix arrays is robust but largely detects cis-acting expression regulators. *Genetics*, **176**, 789–800.

Muller,D.W. and Sawitzki,G. (1991) Excess mass estimates and tests for multimodality. *J. Am. Stat. Assoc.*, **86**, 738–746.

Ronald,J. *et al*. (2005) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res.*, **15**, 284–291.

Rostoks,N. *et al*. (2005) Single-feature polymorphism discovery in the barley transcriptome. *Genome Biol.*, **6**, R54.

Silverman,B.W. (1981) Using kernel density estimates to investigate multimodality. *J. R. Stat. Soc. B*, **43**, 97–99.

Wang,M. *et al*. (2009) Robust detection and genotyping of single feature polymorphisms from gene expression data. *PLoS Comput. Biol.*, **5**, e1000317.

Werner,J.D. *et al*. (2005) Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. *Proc. Natl Acad. Sci. USA*, **102**, 2460–2465.

West,M.A. *et al*. (2006) High-density haplotyping with microarray-based expression and single feature polymorphism markers in arabidopsis. *Genome Res.*, **16**, 787–795.

Winzeler,E.A. *et al*. (1998) Direct allelic variation scanning of the yeast genome. *Science*, **281**, 1194–1197.

Wu,Y. *et al*. (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.*, **4**, e1000212.

Xie,W. *et al*. (2009) Single feature polymorphisms between two rice cultivars detected using a median polish method. *TAG Theor. Appl. Genet.*, **119**, 151–164.