Vol. 21 no. 20 2005, pages 3852–3858
doi:10.1093/bioinformatics/bti640

Gene expression

Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit

Xinping Cui[1,*], Jin Xu[1], Rehana Asghar[3], Pascal Condamine[2], Jan T. Svensson[2], Steve Wanamaker[2], Nils Stein[4], Mikeal Roose[2] and Timothy J. Close[2]

[1]Department of Statistics and [2]Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA, [3]Department of Botany, University of Arid Agriculture, Rawalpindi 46300, Pakistan and [4]Institute of Plant Genetics and Crop Plant Research (IPK), Department Genebank, AG MOM, Corrensstrasse 3, D-06466 Gatersleben, Germany

Received on May 13, 2005; revised on August 17, 2005; accepted on August 18, 2005
Advance Access publication August 23, 2005

ABSTRACT

Motivation: Genomic DNA was hybridized to oligonucleotide microarrays to identify single-feature polymorphisms (SFP) for *Arabidopsis*, which has a genome size of ~130 Mb. However, that method does not work well for organisms such as barley, with a much larger 5200 Mb genome. In the present study, we demonstrate SFP detection using a small number of replicate datasets and complex RNA as a surrogate for barley DNA. To identify single probes defining SFPs in the data, we developed a method using robustified projection pursuit (RPP). This method first evaluates, for each probe set, the overall differentiation of signal intensities between two genotypes and then measures the contribution of the individual probes within the probe set to the overall differentiation.

Results: RNA from whole seedlings with and without dehydration stress provided 'present' calls for ~75% of probe sets. Using triplicated data, among the 5% of 'present' probe sets identified as most likely to contain at least one SFP probe, at least 80% are correctly predicted. This was determined by direct sequencing of PCR amplicons derived from barley genomic DNA. Using a 5 percentile cutoff, we defined 2007 SFP probes contained in 1684 probe sets by combining three parental genotype comparisons: Steptoe versus Morex, Morex versus Barke and Oregon Wolfe Barley Dominant versus Recessive.

Availability: The algorithm is available upon request from the corresponding author.

Contact: xinping.cui@ucr.edu

Supplementary Information: http://faculty.ucr.edu/~xpcui

1 INTRODUCTION

Genetic differences between genotypes include single nucleotide polymorphisms (SNPs) and insertion/deletions (INDELs). Either type of variation, as well as splicing and polyadenylation differences (Rostocks *et al*., 2005), can potentially influence the hybridization of mRNA to 25-mer oligonucleotides. A polymorphism detected by a single probe in an oligonucleotide array is called a single-feature polymorphsim (SFP), where a feature refers to a probe in the array. Polymorphisms within a transcribed sequence are of particular interest because they may reflect variation in biological function. SFPs detected using high density oligonucleotides microarrays such as the Barley1 Affymetrix GeneChip (Close *et al*., 2004) can serve as function-associated markers for genetic analyses including quantitative trait loci (QTL) and linkage disequilibrium (LD) mapping. For yeast (Winzeler *et al*., 1998) and *Arabidopsis* (Borevitz *et al*., 2003) hybridization of total genomic DNA has been a successful route to detecting SFPs. However, the size of the barley genome is ~5200 Mb, which is ~40 times the size of *Arabidopsis*. We found that the DNA hybridization method used by Borevitz *et al*. (2003) is not satisfactory for barley, but instead RNA can serve as a surrogate for DNA. In the current study, we present a statistical method that uses Affymetrix GeneChip data obtained using complex RNA from barley to detect individual probes that reveal SFPs. The idea is based on the expectation that, for each gene, two genotypes will have parallel expression values for all probes in a given probe set, except for probes that cover the SFPs. Recently, Rostocks *et al*. (2005), in an independent study, described an application of RNA-derived barley data using a different statistical approach that considers each probe singularly against all probes on the chip. Also, Ronald *et al*. (2005) used oligonucleotide arrays for genotyping with yeast RNA.

An Affymetrix GeneChip uses a set of probe pairs to measure the expression of each gene. Each probe pair consists of a perfect match (PM) probe and a mismatch (MM) probe. The latter serves the purpose mainly of distinguishing background noise from signal. However, recent studies have shown that MM values also detect hybridization signals (Irizarry *et al*., 2003; Wu *et al*., 2004), which raises questions about the reliability of estimating background noise based on MM values. Therefore, we utilize only the PM values in this analysis. Let S_{tij} be the log-scaled observed PM value of the j-th probe in the i-th probe set hybridized to a RNA sample with genotype t. (Here $t = 1, 2$.) Then it can be modeled as

$$S_{tij} = I_{ti} + A_{tij} + \epsilon_{tij}, \qquad (1)$$

where I represents the expression index at probe set level; A measures the probe affinity effect and ϵ is the random error (Li and Wong, 2001a; Hubbell *et al*., 2002; Wu *et al*., 2004). When the target RNA samples from two different genotypes do not have a sequence variant for a particular gene, the affinity effects are independent of genotypes for all probes that cover this gene. For such cases, attempts have been made to model the affinity effects as a function of the probe sequences synthesized on the GeneChip (Naef and

*To whom correspondence should be addressed.

Magnasco, 2003; Zhang *et al*., 2003; Wu *et al*., 2004). However, when a nucleotide polymorphism (NP) does occur within a gene segment covered by a probe, the affinity effect of that probe changes with the genotype. Therefore, detecting probes with differentiated affinity effect among different genotypes will help locate SFP probes.

Some current models or packages, such as MAS 5.0 (Hubbell *et al*., 2002), dChip (Li and Wong, 2001a), RMA (Irizarry *et al*., 2003), GC-RMA (Wu *et al*., 2004), focus mainly on the estimation of expression index I at the probe set level. However, the study presented here focuses on detecting intensity differences between two genotypes at the probe level while accounting for the overall genotype effect at the probe set level. Although dChip can potentially be used to detect outlying probes which behave substantially differently across genotypes, it usually needs more than 10 biological replicates from each genotype for model training and outlier detection (Li and Wong, 2001b). Rostocks *et al*. (2005) applied a probe level model based on the Bioconductor package siggenes and statistical analysis of microarrays (SAM) to identify SFPs between 17 datasets of one genotype and 19 of another genotype. In that work, a subset of ~4% of all perfect match probes was highly enriched for valid SFPs since it included 67% of known SNPs. In the present study, we present an alternative method that requires as few as three biological replicates to identify within a smaller subset of the data certain types of SFP probes that are particularly useful. It is based on the method of robustified projection pursuit. We first use it to calculate an outlying score for every probe set with 'present' call. A large value suggests that the probe set may contain SFP probe(s). Second, we further evaluate the contribution of each probe to the overall outlyingness of the probe set and then locate the responsible SFP probe(s). The algorithm was trained on polymorphisms apparent in the HarvEST:Barley EST database (Wanamaker and Close, 2004, http://harvest.ucr.edu). The validation rate of a priori SFP detection from GeneChip data was measured by PCR amplification and amplicon sequencing and found to be >80% when only probe sets with 'present' calls are considered.

2 EXPERIMENTAL DESIGN

We produced data using the Affymetrix Barley1 GeneChip (Close *et al*., 2004) hybridized with cRNA synthesized from young seedling RNA of five genotypes: Steptoe, Morex, Barke, Oregon Wolfe Barley (OWB) dominant and OWB recessive. These genotypes were chosen because they are the parents of three doubled haploid (DH) mapping populations. The Steptoe × Morex map presently contains the largest number of markers from any barley mapping population (http://wheat.pw.usda.gov/ggpages/SxM/). The Oregon Wolfe barleys (http://www.barleyworld.org) are more polymorphic than Steptoe × Morex and have recently become one of the most preferred mapping materials. The Morex × Barke DH population was developed most recently and is attractive in part because Barke and Morex have been the No.1 and No.2 sources of public EST sequences, respectively (http://www.ncbi.nlm.nih.gov/dbEST/). In addition, a BAC library from Morex barley (Yu *et al*., 2000) is widely used.

Analysis of more than 40 wheat cDNA libraries revealed that drought-stress greatly increased the complexity of RNA populations in seedlings (Zhang *et al*., 2004). We verified with HarvEST:Barley (http://harvest.ucr.edu) that this is also true for barley. We made use of this observation to create highly complex RNA populations, as a surrogate for DNA, from each genotype by pooling equal quantities of RNA from stressed and unstressed whole seedlings for Steptoe, Morex and Barke. For the Oregon Wolfe Barley samples, stressed and unstressed whole-seedling RNA samples were analyzed separately. Briefly, seeds were surface sterilized, then germinated in glass crystallization dishes on wet filter paper in the dark until roots emerged and shoots were ~5 cm long. Unstressed etiolated samples were collected by scraping out the softened starchy endosperm with a spatula, then pulverizing the entire remaining tissue for total RNA and protein extraction. Five seedlings were pooled for each sample. The same type of seedling tissues were harvested also after placing seedlings in a 90% relative humidity chamber for 48 h under sufficient lighting to evoke greening of the shoots. The expression of typical drought-stress proteins was verified using anti-dehydrin antibodies (Close *et al*., 1993) for western blot analysis. RNA concentrations were determined by UV spectroscopy and RNA integrity was verified according to standard methods. For each of the five genotypes, the preparation of stressed and unstressed seedling RNA was done three times, each a temporally independent replicate. The 50:50 mixtures of stressed and unstressed RNA typically yielded 'present' call rate in the range of 70–80% of all probe sets for each RNA sample [consult Affymetrix manual (2001) for the details of 'present' calls]. For SFP detection, we used only the probe sets with 'present' calls in all six samples from each two genotypes under comparison to enhance the signal-to-noise ratio. This resulted in ~14 000 (out of 22 801) probe sets used in each comparison. The array data have been deposited into the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/, GSE3170).

3 METHODS AND ALGORITHM

We assume that $\text{median}_j\{A_{tij}\} = 0$ and ϵ_{tij} is symmetric (Hubbell *et al*., 2002; Li and Wong, 2001) so that model (1) is identifiable. Then for the i-th probe set, under least absolute error, we have the estimators of the expression index and the affinity effect given by

$$\begin{cases} \hat{I}_{ti} = \text{median}_j\{S_{tij}\}, \\ \hat{A}_{tij} = S_{tij} - \hat{I}_{ti}. \end{cases} \tag{2}$$

Denote the difference of affinity effect between two genotypes by Y, a $N \times p$ matrix, which is given by

$$Y_{ij} = \text{sample median of } \hat{A}_{1ij} - \text{sample median of } \hat{A}_{2ij}, \tag{3}$$

where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, p$. More specifically, N row vectors of Y represent N distinct probe sets after filtering and p column vectors represent p probes which are tiled across a gene. We denote N row vectors by $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$ for future use. Here the sample median is taken on the three biological replicates from each genotype.

In many cases we can see the parallelism of signal intensity between two genotypes for every probe in a probe set. Such constant differentiation between p perfect match probes is owing to genotype dependent expression polymorphism. The fluctuation of signal intensities of probes in the same probe set reflects difference in hybridization affinity among probes that cover different sequence segment of the same gene. Such fluctuation remains the same among different genotypes unless a polymorphism covered by a probe

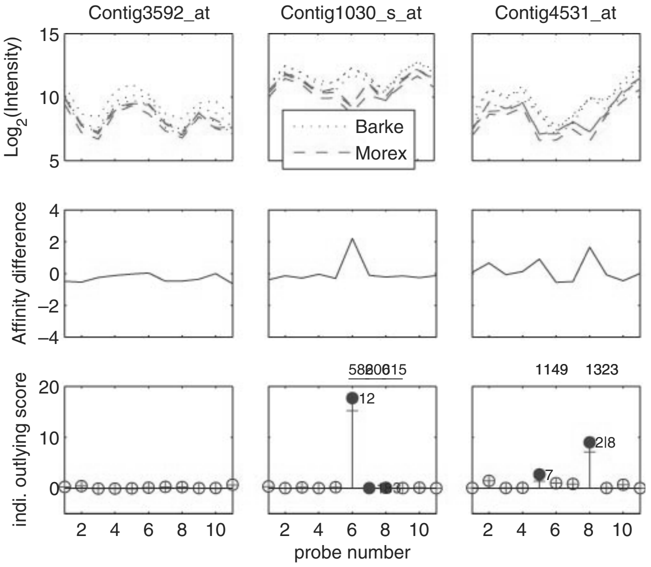

Fig. 1. (Top row) Plots of the log intensities for three probe sets from two genotypes. (Middle row) Plots of the differentiations of average log intensities between two genotypes adjusted for genotype dependent expression differences. (Bottom row) Plots of individual outlying scores, where 'plus' represents w_{ij}, 'open circles' represent w^*_{ij}, 'closed circles' indicate the actual occurrence of a EST validated SNP (see text for definitions of w_{ij} andw^*_{ij}). The SNP position within the probe is noted beside the bullet, and multiple positions are delimited by vertical bars. For example, probe 8 of Contig4531_at has 2 SNPs at positions 2 and 8. The starting positions (with respect to the consensus gene) of probes with SNPs are shown as well. Their numbers are underlined if the corresponding neighboring probes overlap, i.e. the starting positions are within 25 nucleotides.

exists. For instance, in the middle row of Figure 1, Contig3592_at (probe set name) has its differentiation ~0 (after adjusting for genotype-dependent expression polymorphism); Contig1030_s_at and Contig4531_at have most intensity differentiations ~0 but shifts at probe 6 and probes 5 and 8, respectively. These shifts, either upward or downward, indicate with high probability the occurrence of SFPs, although it can also be possible that the unusual intensity differentiations may be attributed to measurement error. The likelihood of measurement error, of course, decreases as the number of replicates increases.

Since we expect that the majority of nucleotide positions within genes in the barley genome do not have polymorphism, parallel pattern of signal intensity between two genotypes should be observed in most probe sets. From a geometric point of view, if differentiation in signal intensity between two genotypes is represented by a p-dimensional point for each probe set, they would form a cloud in p-dimensional space with the majority of points clustered together and any point at the edge suggests a 'potential' probe set that might contain SFP probe(s). Projection pursuit can provide a robust measure of the outlyingness of an observation in a sample (Rousseeuw and Leroy, 1987). Here we use it to calculate overall outlying scores and individual outlying scores (defined below) to separate 'potential' probe sets from the whole collection of probe sets under consideration and to quantify the contribution of individual probes to the overall outlyingness of their affiliated probe sets.

The algorithm can be summarized as follows.

(1) Fix a direction $\boldsymbol{\nu}$ (a $p \times 1$ vector). Project Y onto $\boldsymbol{\nu}$.

(2) Use relative absolute deviation to measure the outlyingness for every probe set on $\boldsymbol{\nu}$.

(3) Repeat steps (1) and (2) for all directions and take the supremum as the final overall outlying score for the probe set.

Let u_i be the overall outlying score for probe set $\boldsymbol{y}_i$, $i = 1, 2, \dots, N$. It is then defined as

$$u_i = u_i(Y,\ \text{all}\ \boldsymbol{\nu}) = \sup_{\text{all}\,\boldsymbol{\nu}} \frac{|\boldsymbol{y}_i'\boldsymbol{\nu} - \text{med}_j\,(\boldsymbol{y}'_j\boldsymbol{\nu})|}{\text{med}_k|\boldsymbol{y}'_k\boldsymbol{\nu} - \text{med}_j(\boldsymbol{y}'_j\boldsymbol{\nu})|}, \qquad (4)$$

where $\boldsymbol{y}'_j\boldsymbol{\nu}$ is the usual inner product, i.e. $\boldsymbol{a}'\boldsymbol{b} = a_1b_1 + a_2b_2 + \cdots + a_pb_p$; med stands for the median (Rousseeuw and Leroy, 1987).

The optimal properties of this measurement are well known. First, it is robust. We use the median and median absolute deviation to measure the center and the variation for the projected data rather than the mean and the SD whose values are very sensitive to outliers in the sample. Second, u_i is affine equivariant, that is,

$$u_i(Y) = u_i(C\boldsymbol{y}_h + \boldsymbol{d}), \quad h = 1, 2, \dots, N, \qquad (5)$$

for any non-singular $p \times p$ matrix C and any vector $\boldsymbol{d}$ in $\mathbb{R}^{\mathrm{p}}$. In other words, this statistic captures similarity in 'shape' of signal intensities among probe sets, but places no emphasis on the magnitude of intensities.

In practice, we cannot try all directions in the p-dimensional space since there are infinitely many. It is suggested to use N directions represented by the observations (Rousseeuw and van Zomeren, 1990). However we find it is unnecessary in our case since most of the probe sets have small differentiation between two compared genotypes, which is around the baseline zero after adjustment. We suggest using only those row vectors having high variation. This work can be done by some preliminary selection. We chose ~2000 directions. As it turned out, the final result did not change much when we added more directions.

Next, we evaluate the individual contribution by each probe in a probe set. It is desired that the individual outlying score should reflect the corresponding differentiation around baseline. Let w_{ij} be the individual outlying score of the j-th probe in the i-th probe set. We expect w_{ij} to be small if the differentiation is close to the baseline and a positive number proportional to the magnitude of the shift if a polymorphism exists. We first propose the following definition,

$$w_{ij} = u_i(Y) - u_i\big(Y \text{ with } Y_{ij} \text{ replaced by } \text{med}_j\{Y_{ij}\}\big), \qquad (6)$$

where u_i is defined in Equation (4). This definition is based on the observation that the second term in the RHS of Equation (6) is a measure of overall outlying score of the i-th probe set after replacing the effect of the j-th probe by the baseline which is zero after the adjustment. It is expected to be no greater than $u_i(Y)$ so that the difference between two overall outlying scores can be treated as the j-th probe's contribution to the total outlyingness.

In practice, since we use finite projection directions, u_i is actually obtained at one particular direction, say $\boldsymbol{\nu}^*_i$, along which the relative absolute deviation is maximized. Then it is more appropriate to compute the overall outlying score for the projected data along $\boldsymbol{\nu}^*_i$ alone. Hence a modification for w_{ij} is given by

$$w^*_{ij} = u_i\big(Y, \boldsymbol{\nu}^*_i\big) - u_i\big(Y \text{ with } Y_{ij} \ \text{ replaced by } \text{med}_j\{Y_{ij}\}, \boldsymbol{\nu}^*_i\big). \quad (7)$$

Obviously, we have $w^*_{ij} \geqslant w_{ij}$ (Fig. 1). The fact that w^*_{ij} amplifies on the differentiation more than w_{ij} is appealing since it increases the sensitivity of SFP detection. Another advantage is that w^*_{ij} only needs to be computed for one direction in the second term.

At last, the probe sets with the highest overall outlying scores will be identified as containing putative SFP probe(s). Then an SFP will be located at the probe with the highest individual outlying score. When multiple SFPs are involved, one can define the selection rule with certain stringency according to the real situation.

4 RESULTS

4.1 SFP detection

Applying the method to the Barley1 array data, we obtain u_i, w_{ij} and w^*_{ij} for N probe sets. The distribution of u in the Barke and Morex comparison is shown by the histogram in Figure 2. The candidate probe sets are selected from the right tail of this distribution, which corresponds to high outlying scores. The actual cutoff can be defined by the user based on the distribution of u. We see in the bottom row of Figure 1 an example where our findings of SFPs match well with SNPs supported by the EST database. Table 1 shows the number of SFPs probes detected at different thresholds for the three genotype combinations (see also supplementary Table).

To measure the proportion of correct SFP calls in triplicated datasets, we chose a random sample of 64 probe sets containing 65 putative SFP probes for direct sequencing validation. These 64 probe sets were from the analyses of Morex versus Steptoe or (the union, not the intersection) OWB Dominant versus OWB Recessive using the top 5 percentile as the cutoff. Of these four genotypes, EST sequence information existed only for Morex, so PCR amplification of genomic DNA and direct sequencing were carried out on all four genotypes. A total of 52 (80%) of these 65 SFP probes were found to cover polymorphisms. Among these, 28 (54%) were positioned over a single SNP, 12 (23%) over >1 SNP, 11 (21%) over an indel and 1 (2%) over an SNP and an indel. It is also noteworthy that 7 of the remaining 13 probes belonged to probe sets that correspond to multigene families. The fact that we may have chosen the wrong family member for sequence validation argues that the 20% false discovery rate determined by direct sequencing is likely to be an overestimate, and thus 80% must be a conservative estimate of the percentage of correct SFP calls when we consider only probe sets with 'present' calls. Therefore, we claim that at the threshold of top 5 percentile, our detection method is correct at least 80% of the time in a priori identification of SFPs on a filtered set of ~14 000 'present' probe sets from the Barley1 GeneChip using triplicate RNA-based datasets.

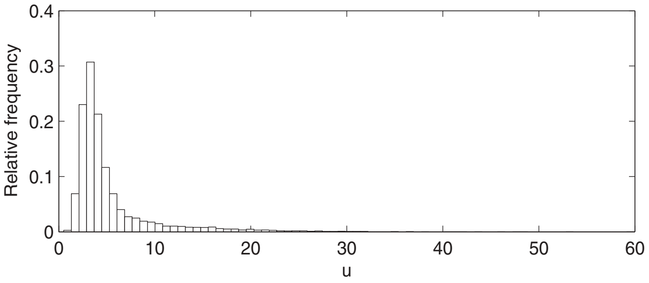

Fig. 2. Histogram of the relative frequencies of the overall outlying scores (u) of probe sets in the Morex versus Barke dataset.

Table 1. SFP probes called from three parental genotype comparisons

Top percentile (%)	BM	SM	OWB
25	3635	3626	4832
20	2918	2907	3794
15	2200	2190	2833
10	1480	1474	1911
5	745	740	986

BM, Barke versus Morex; SM, Steptoe versus Morex; OWB, Oregon Wolf Barley Dominant versus Recessive.

A second source of information on the efficacy of our method comes from the database of NPs indicated by EST-based unigene sequences (eNPs). We analyzed Morex and Barke EST sequences present in the HarvEST:Barley assembly #32 (http://harvest.ucr.edu). This assembly contains 2904 unigenes having at least two Morex and two Barke EST sequences that overlap at least one 25-mer probe. Among these unigenes, 385 contain at least one probe that covers a Morex versus Barke eNP, a total of 696 probes from 391 probe sets. We determined that the Morex versus Barke eNPs have a validation rate of ~90.6% (29/32) by direct sequencing of Morex and Barke amplicons. We first narrowed the list of eNPs to only those that were covered by 25-mer probes whose probe sets had present calls in all six Morex and Barke datasets. This reduced the number of eNPs for further consideration to 510 probes from 319 probe sets. These 510 probes cover 377 unique eNPs, We then asked what portion of these 'known' 377 eNPs were detected by our method. Figure 3 shows the sensitivity of detection of eNPs at different probe set outlying score thresholds. The solid line in Figure 3A demonstrates the eNP detection rates when considering all 377 eNPs. Here a probe in a detected probe set will be called an SFP probe if it contributes >40% of the overall outlying score. The dashed line (Fig. 3A) shows the eNP detection rates when considering only eNPs located in the central region of the 25-mer (positions 6–20). The curves in Figure 3A have limits much <100% because we consider only the probes contributing at least a fairly high minimal portion of the overall outlyingness of their detected probe set. For example, this method often discards multiple probes that are tiled over the same SNP. In addition, as discussed in Section 4.3, probes having an SNP at the boundaries are unlikely to be detected. Therefore, although increasing the threshold will increase the number of identified probe sets, the proportion of eNPs detected will reach a limit, which in this case is 47.7% and 61.1% for solid and dashed lines. An alternative representation of the approach to this limit is shown in Figure 3B, where 76.1% of the maximum RPP detectable eNPs and 82.6% of the maximum RPP detectable centrally located eNPs can be obtained using the 15 percentile threshold. This suggests that the 15 percentile is a reasonable threshold value in this case. However, the validation rate for this larger subset of the data may be less than for the 5 percentile threshold, as discussed later.

Knowledge of genomic sequences is not sufficiently complete from the available EST dataset to unequivocally define a set of known NPs or sequences known not to contain NPs. As noted above, we found that ~90.6% of predicted eNPs are correct. However, the method that we used to achieve such a high eNP validation

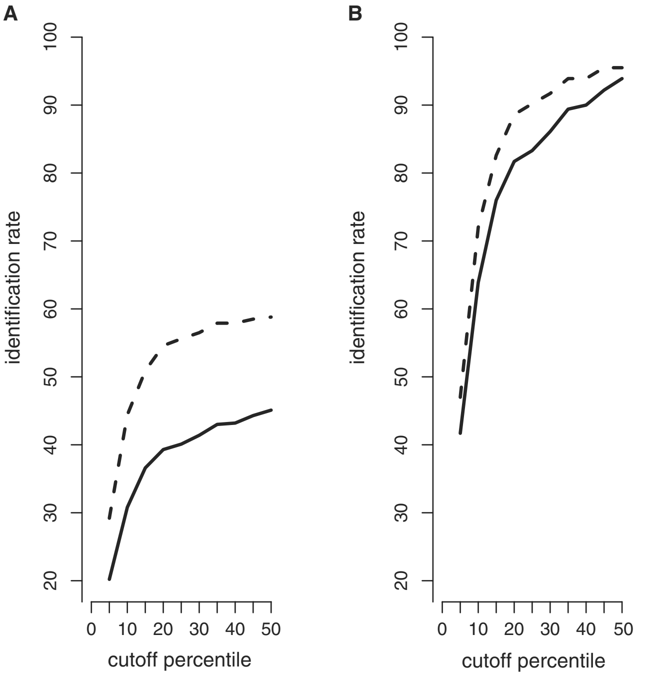

Fig. 3. Cutoff threshold versus sensitivity of detecting (**A**) SFP probes when considering only probes contributing >40% of overall outlying scores (solid line); SFPs located in the central region of 25-mer (dashed line); (**B**) maximum RPP identifiable SFPs (solid line); maximum RPP identifiable centrally located SFPs (dashed line).

rate discards the majority of actual NPs simply for lack of sufficiently compelling EST data. Therefore, the eNPs derived from the EST dataset cannot serve as a satisfactory component of calculations for false discovery rate.

4.2 Single SFP probe versus multiple SFP probes

Based on the eNP dataset, it is found that the overall outlying score of a probe set containing multiple SFP probes is stochastically greater than that of a probe set containing only one SFP probe, i.e.

$$\text{prob}(u(\text{multiple SFPs}) > u(\text{single SFP})) > \tfrac{1}{2}. \tag{8}$$

The distributions seem to share the same pattern except for a location shift or scale difference (Fig. 4).

4.3 Individual outlying score against the position

Since we know from EST sequences the exact SNP positions within many probes, it is of interest to see the relationship between the individual outlying score and the SNP position within the probe. For simplicity, we show the data only for the probes that contain only one SNP. From Figure 5, we see that the probes having an SNP around the center are more likely to yield a higher individual outlying score than those at the boundaries. Probe sets sometimes contain multiple SFP probes tiled over a single SNP. In most of these cases the SNP position effect is consistent with this overall behavior. An example is shown in Figure 6 where the 25-mer probes that position the SNP more centrally have the higher individual outlying scores.

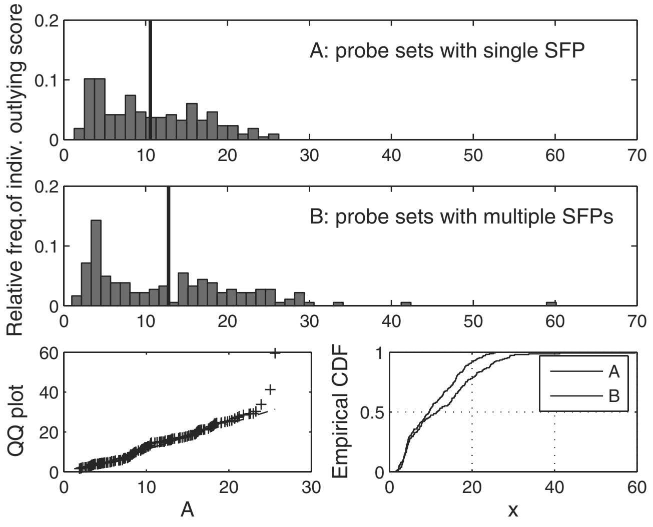

Fig. 4. Histogram of u for probesets containing single or multiple SFP probes. The Q–Q plot shows u (probe set with multiple SFPs) and u (probe set with a single SFP) belong to the same distribution except for a location or scale difference. Two sample means are noted by vertical bars in the histograms. The fact that the empirical cumulative distribution function (CDF) of A is entirely over that of B indicates the relationship in Equation (8).

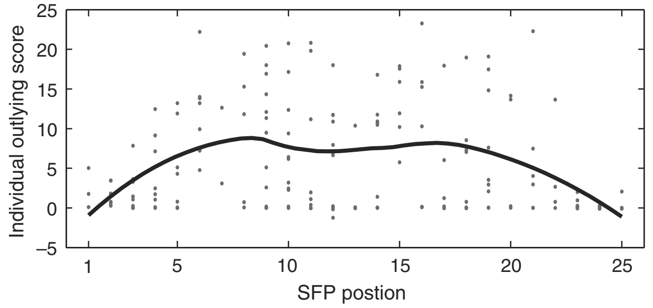

Fig. 5. Scatter plot of individual outlying scores against the SFP positions within the corresponding probes. A smoothing curve by Loess method is added.

Also, the predicted SFP probes were divided into three categories and compared with eNPs defined from the available Morex and Barke EST sequences (Table 2). The three categories were single eNPs between positions 6 and 20 in the 25-mer, single eNPs positioned outside this 6–20 region, and multiple eNPs residing in one probe. As shown in Table 2, SFP probe sets in the top 15 percentile identify ∼51% (110 of 216) of the EST-supported eNPs located within the range of positions 6–20, whereas only 15.5% of eNPs (16 of 103) located outside this region are identified. This also suggests that polymorphisms residing in the central region of the 25-mer have higher probabilities of being detected by the RPP method. This phenomenon has also been reported elsewhere (Borevitz *et al*., 2003; Rostocks *et al*., 2005). Table 2 also shows that the top 5 percentile probe sets include ∼29% of centrally-positioned eNPs, but only ∼3% of eNPs that are located less centrally. This indicates that the 5 percentile threshold is considerably more selective for centrally positioned NPs than is the 15 percentile threshold.

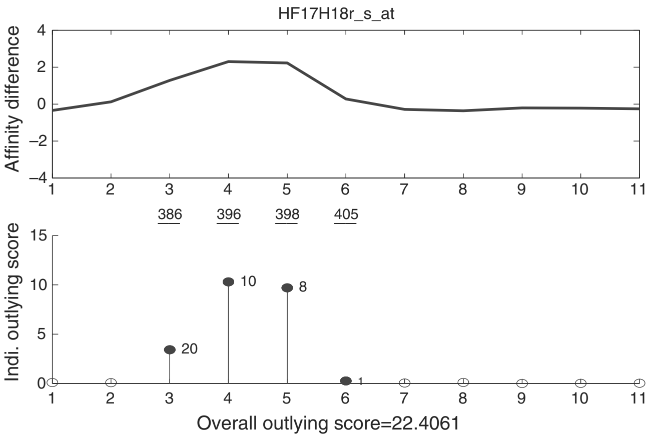

Fig. 6. An example of multiple probes that overlap the same SNP. The probes (4 and 5) having an SNP around the center (positions 10 and 8) yield higher individual outlying scores than those (positions 3 and 6) having an SNP at the boundaries (positions 20 and 1).

Table 2. RPP-predicted SFPs compared with EST-supported eNPs

	377 EST-supported eNPs among 319 probe sets		
Top percentile (%)	216 single eNPs $\subsetneq$ [6–20mer]	103 single eNPs $\nsubseteq$ [6–20mer]	58 multiple eNPs
25	120	18	13
20	118	17	13
15	110	16	12
10	96	9	11
5	63	3	10

5 DISCUSSION

We have presented a statistical approach based on robustified projection pursuit to identify SFPs using Affymetrix GeneChip expression data. Each SFP was defined by a comparison of two genotypes with three biological replicates for each genotype. We used log transformed PM values without background correction and normalization. Although background correction and quantile normalization proposed by Irizarry *et al*. (2003) have been widely used, not applying background correction mainly causes a problem for small intensity values which are known to be very noisy. As pointed out by Irizarry *et al*. (2003), a significant problem with quantile normalization is the risk of removing some of the signal in the tail. Since we used GCOS 1.2 ‘present call’ to filter out small, noisy intensity values and we are very interested in signal in the tail for SFP detection, we wanted to avoid this known pitfall of background correction and quantile normalization. As a comparison, we applied the RMA background correction, quantile normalization and log transformation suggested by Irizarry *et al*. (2003) and used by Rostocks *et al*. (2005) to our raw PM values and then repeated RPP analysis; we found that the rate of detected SFP probes matching with eNP list uniformly decreased at different threshold cutoff (data not shown). Theoretically speaking, however, background correction and normalization should help find real SFPs since they will remove some noise from data and therefore they help reduce the number of falsely called SFP probes. Since the existing background correction and normalization method do not seem to be very effective based on our analysis, developing a new data denoising technique will be our future research.

We can see from model (3) the differentiation measurements y between two genotypes have been adjusted for overall gene-expression level of each genotype. Therefore, any probe-level intensity differentiation between two genotypes indicates a potential SFP. However, instead of directly analyzing significance of differentiation at each probe independently (Rostocks *et al*., 2005), our method takes into account the correlation between probes in the same probe set by first evaluating significance of overall differentiation for each probe set.

We envision some specific circumstances where our method of SFP detection will be particularly relevant. For example, researchers often wish to have one thousand to a few thousand markers for the development of a genetic map from mapping populations that range in size generally from one hundred to several thousand individuals. Using our method, one can readily generate a list of SFP markers from two reasonably polymorphic parental genotypes using as few as three replicate GeneChip datasets from each parent. It would be possible then to either use the same GeneChip that was used to define the SFPs to genotype each member of the mapping population or develop a new, less costly chip containing only the probe sets that contain SFPs. Since our method identifies individual SFP probes in the context of the probe set rather than in the context of all individual probes on a chip, the data from a smaller chip would have no less genotyping power than the original, larger chip. In addition, we have observed that the percent false SFPs at the 5 percentile threshold is not significantly different when using triplicated or duplicated datasets. As stated above, the false SFP rate was <20% (13/65) with a triplicated dataset from Morex and Barke; it was virtually the same error rate (18%;13/73) with duplicated datasets from the same two genotypes. Based on this observation, and considering that there will be an additional level of dataset duplication among individuals in mapping populations, we suggest a cost-efficient SFP mapping path that involves (1) three replicates of each parent using complex RNA, (2) selection of a percentile cutoff between 5 and 15% to create a RPP-based SFP list of the desired length, (3) two replicates of each genotype in the mapping population and possibly (4) the production of a smaller chip, depending on the population size and other cost-efficiency factors. Similarly, a list of diversity-related SFPs can be created by compiling RPP-based SFPs from various genotype combination to extend SFP marker analysis to germplasm accessions from within a species, again with the option of creating a smaller chip than the full chip used for initial RPP-based SFP discovery. Smaller chips from barley or any species presumably soon can be placed in 96-well format, further reducing the cost per chip and the need for human labor. The identification of RPP-compatible probe sets bearing SFP probes support these future developments.

ACKNOWLEDGEMENTS

We wish to thank Kavitha Kothakonda and Matthew Lyon for the helpful discussion. We also wish to thank the two anonymous referees, whose demands for rigor and clarity helped us improve our paper. This work was supported in part by UCR AES-CE RSAP

A01869, 'Statistical Design and the Analysis of Gene Expression Microarray Data', NSF DBI-0321756, 'Coupling Expressed Sequences and Bacterial Artificial Chromosome Resources to Access the Barley Genome' and USDA-NRI 02-35300-12548, 'HarvEST: A Portable EST Database Viewer'.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2001), Statistical algorithms reference guide. *Technical Report*, Affymetrix.

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

Close,T.J. *et al*. (1993) A view of plant dehydrins using antibodies specific to the carboxy terminal peptide. *Plant Mol. Biol*., **23**, 279–286.

Close,T.J. *et al*. (2004) A new resource for cereal genomics: 22K Barley GeneChip comes of age. *Plant Physiol*., **134**, 960–968.

Hubbell,E. *et al*. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.

Irizarry,R.A. *et al*. (2003) Summary of Affymetrix GeneChip probe level data. *Nucleic Acids Res*., **31**, e15.

Li,C. and Wong,H.W. (2001a) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.

Li,C. and Wong,H.W. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*, **2**, research0032.1-0032.11.

Naef,F. and Magnasco,M.O. (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Phys. Rev. E*, **68**, 011906.

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

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

Rousseeuw,P.J. and Leroy,A.M. (1987) *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.

Rousseeuw,P.J. and van Zomeren,B.C. (1990) Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc*., **85**, 633–639.

Wanamaker,S. and Close,T.J. (2004) HarvEST:Barley version 1.32.

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

Wu,Z., Irizarry,R.A., Gentleman,R., Murillo,F.M. and Spencer,F. (2004) A model based background adjustment for oligonucleotide expression data. *J. Am. Stat. Assoc*., **99**, (468), 909–917.

Yu,Y. *et al*. (2000) A bacterial artificial chromosome library for barley (*Hordeum vulgare L.*) and the identification of clones containing putative resistance genes. *Theor. Appl. Genet*., **101**, 1093–1099.

Zhang,L. *et al*. (2003) A model of molecular interactions on short oligonucleotide microarrays. *Nat. Biotechnol*., **21**, 818–821.

Zhang,D. *et al*. (2004) Construction and evaluation of cDNA libraries for large-scale expressed sequence tag sequencing in wheat (*Triticum aestivum L.*). *Genetics*, **168**, 595–608.