Theor Appl Genet (2009) 119:151–164
DOI 10.1007/s00122-009-1025-2

ORIGINAL PAPER

Single feature polymorphisms between two rice cultivars detected using a median polish method

Weibo Xie · Ying Chen · Gang Zhou · Lei Wang · Chengjun Zhang · Jianwei Zhang · Jinghua Xiao · Tong Zhu · Qifa Zhang

Received: 20 March 2009 / Accepted: 23 March 2009 / Published online: 16 April 2009

Abstract Expression levels measured in microarrays of oligonucleotide probes have now been adapted as a high throughput approach for identifying DNA sequence variation between genotypes, referred to as single feature polymorphisms (SFPs). Although there have been increasing interests in this approach, there is still need for improving the algorithm in order to achieve high sensitivity and specificity especially with complex genome and large datasets, while maintaining optimal computational performance. We obtained microarray datasets for expression profiles of two rice cultivars and adapted a median polish method to detect SFPs. The analysis identified 6,655 SFPs between two the rice varieties representing 3,131 rice unique genes. We showed that the median polish method has the advantage of avoiding fitting complex linear models thus can be used to analyze complex transcriptome datasets like the ones in this study. The method is also superior in sensitivity, accuracy and computing time requirement compared with two previously used methods. A comparison with data from a resequencing project indicated that 75.6% of the SFPs had SNP supports in the probe regions. Further comparison revealed that SNPs in sequences immediately flanking the probes also had contributions to the detection of SFPs in cases where the probes and the targets had perfectly matched sequences. It was shown that differences in minimum free energies caused by flanking SNPs, which may change the stability of RNA secondary structure, may partly explain the SFPs as detected. These SFPs may facilitate gene discovery in future studies.

Communicated by A. Melchinger.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-009-1025-2) contains supplementary material, which is available to authorized users.

W. Xie · Y. Chen · G. Zhou · L. Wang · C. Zhang · J. Zhang · J. Xiao · T. Zhu · Q. Zhang (✉)
National Key Laboratory of Crop Genetic Improvement and National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, 430070 Wuhan, China
e-mail: qifazh@mail.hzau.edu.cn

Introduction

Short DNA oligonucleotides are commonly used as probes to interrogate nucleic acid targets in microarray analyses. These oligonucleotide probes, as short as 25 bases, are used in one of the most widely used microarray technology, Affymetrix GeneChip microarray. The short probes are generally more sensitive to mismatch bases to the targets during molecular hybridization comparing to probes with longer sequences. It is known that a mismatch base within the probe sequence could reduce the binding affinity between targets and probes (Borevitz et al. 2003, 2007; Ronald et al. 2005; Zhu and Salmeron 2007).

The feature about the high sensitivity to sequence variations of the short oligonucleotide probes has two implications in the microarray analysis: (1) in order to accurately measure transcript abundance, perfect match is essential between probes and targets and between targets from different samples within the probe; (2) if target abundance is presumed to be at similar level in different samples, difference in hybridization signals measured by the same probe sequence could be used to infer sequence mismatch between targets from different samples. The latter has been developed into an approach to quickly identify genetic variations, referred to as single feature polymorphisms (SFPs), by genomic DNA hybridization

with microarrays in a number of species, including yeast, *Arabidopsis* and rice (Borevitz et al. 2003; Kumar et al. 2007; Winzeler et al. 1998). SFPs have also been detected in genomes with larger size and higher complexity from labeled RNA transcript derivatives (Cui et al. 2005; Das et al. 2008; Luo et al. 2007; Rostoks et al. 2005). Since oligonucleotide microarrays are composed of millions of probes representing most of the expressed unique sequences in the genome, SFPs can capture large number of sequence variations. Thus chip-based SFP detection can serve as high-throughput platforms for genetic analyses, providing thousands of markers in a single experiment (Das et al. 2008; Ronald et al. 2005), which is much more efficient than other molecular markers such as RFLPs and SSRs that have been widely used in genetic analyses. SFP markers have been used for high-resolution genetic mapping (West et al. 2006), and for studying expression quantitative trait loci (eQTLs) (Luo et al. 2007; Potokina et al. 2008; West et al. 2006, 2007).

SFPs have been generally regarded as resulting from mismatched bases between the targets and probes. However, results from previous studies indicated that SFPs can also be repeatedly detected even though the targets and the probes have perfectly matched sequences (Borevitz et al. 2007; Luo et al. 2007; Zhu and Salmeron 2007). Such results have frequently been ignored in the analyses, thus the cause has not been investigated. A possible cause for the SFPs resulting from perfectly matched probe-target sequences might be the influence of polymorphisms, mostly SNPs, between the targets in the sequences flanking the probes (probe flanking SNPs or flanking SNPs).

A highly efficient algorithm is critical for sensitivity and specificity of SFP detection, while maintaining optimal computational performance, especially with complex genome and large data sets. Several algorithms have been developed to detect SFPs from RNA derived microarray data (Cui et al. 2005; Luo et al. 2007; Ronald et al. 2005; Rostoks et al. 2005). These algorithms have to reduce the perturbation of hybridization signals resulting from different transcript abundance rather than from genetic variations to ensure sufficient detection power and accuracy. For example, Rostoks et al. provided a mean-based analysis of variance (ANOVA) method employing a complex linear model for the specific experimental designs, which successfully discriminated the differential hybridization signals inducted by different tissues and genotypes (Rostoks et al. 2005). Ronald et al. inferred SFPs by comparing the observed hybridization signal with the expected using the positional-dependent nearest-neighbour (PDNN) model which needs to estimate 82 energy parameters for specific and nonspecific RNA–DNA binding relying on information of probe sequences (Luo et al. 2007; Ronald et al. 2005). Since these algorithms involved estimation of multiple parameters and were developed for and validated using data from a single organism under study, the efficiency and generality should be evaluated using multiple species with different genome and tissue complexity.

Median polish is a median-based data analysis technique which is considered to be more robust than ANOVA (Seheult and Tukey 2001), thus may provide an alternative approach to dissecting the hybridization signal into transcript abundance and genetic variation. This method has the advantage of avoiding fitting complex linear models thus can be used to analyze the complex transcriptome data sets. In this study, we analyzed SFPs between two rice cultivars based on a microarray datasets for expression profiles using the median polish method. We showed that, in addition to simplicity, this method also has better detection sensitivity and accuracy than two previously used methods. We identified 6,655 SFPs between two rice varieties representing 3,131 rice unique genes. A comparison with data from a resequencing project indicated that 75.6% of the SFPs had SNPs in the probe regions. We also showed that SNPs in sequences flanking the probes also had contributions to the detection of SFPs in cases where the probes and the target sequences had perfect matches, demonstrating the importance of the probe flanking SNPs in comparative transcriptome analyses using oligonucleotide microarrays.

Materials and methods

Plant materials, RNA isolation and microarray hybridizations

Two rice varieties of indica subspecies, Minghui 63 and Zhenshan 97, were used in this study. The seeds were planted and seedlings transplanted in the Experimental Farm of Huazhong Agricultural University, Wuhan, China at three time points to form three biological repeats. Panicle samples were collected at three developmental stages: secondary branch primordium differentiation (SBP), pistil/stamen primordium differentiation (PSP), and pollen-mother cell formation (PCF). The details of the developmental stages and sampling were as described previously (Huang et al. 2006). Approximately, 100 mg of sample tissues was collected under a dissecting microscope at 8:00 am–9:30 am, put into a 1.5-ml-microfuge tube with 1.0 ml TRIZOL (Qiagen), and quickly stored in liquid nitrogen. RNA isolation, labeling, and microarray hybridization were performed by the GeneTech Biotechnology Limited Company (Shanghai, China) according to Affymetrix standard protocols. Six replicates of labeled

RNA targets, representing three biological replicates and two technical replicates, were hybridized to the Rice GeneChip Genome Array (http://www.affymetrix.com/products/arrays/specific/rice.affx). This array contains 57,381 probe sets representing 51,279 transcripts from japonica and indica. Each probe set consists of 11 pairs of 25-mer perfect match (PM) and mismatch (MM) probes which differ only at the middle base.

Additional microarray data

In addition to the rice microarray data generated from this study, microarray data used for yeast allelic specific expression analysis (Ronald et al. 2005) and for barley SFP detection (Rostoks et al. 2005) were also used for the analysis. Yeast data generated by Affymetrix YGS98 GeneChip array were downloaded from NCBI Gene Expression Omnibus (GEO) (acc: GSE1975) (Barrett et al. 2007) and barley data generated by Affymetrix Barley1 GeneChip array can be downloaded from author's website (http://naturalsystems.uchicago.edu/naturalvariation/barley/).

Data processing and SFP detection

The CEL files containing raw intensity data for each probe were read into a R program (Ihaka and Gentleman 1996). Background correction and quantile normalization were performed using RMA methods in Bioconductor affy package (Gautier et al. 2004; Gentleman et al. 2004). The gene expression presence/absence detection (absolute call) was calculated from the CEL files using the Microarray Analysis Suite (MAS 5.0, Affymetrix). Only probes from probe sets with "Present" value in the absolute calls in at least four out of total six arrays of each sample were selected for further analysis. Matrices containing the background corrected and normalized data with "Present" calls by combining the information from the above processing procedures were subjected to median polish analysis to extract the residuals. The limma package (Smyth 2004) was then used to determine SFPs by a selected threshold based on Benjamini and Hochberg (BH) adjusted P values. The custom script and microarray data can be downloaded in our website (http://www.ncpgr.cn/supplements/sfp/).

The method of fitting linear model proposed by Rostoks et al. (2005) was also used for comparing the efficiency of SFP detection. Probes in a set were fit with the following linear model for rice and barley data:

$$\log\left(Y_{tgrp}\right) = \mu + t + g + g \times t + p + \varepsilon$$

where Y is the background corrected normalized intensity of t (tissue), g (genotype), r (replicate), and p (probe) in a probe set; μ the mean probe intensity. For yeast data set, ploidy (haploid or diploid) was treated as the tissue component in the plant data set. The residuals which embodied the genotype × probe effects were thus extracted to identify SFPs. The scripts and data of this method can be downloaded from authors' web site (see "Additional microarray data").

Rice probe sequence mapping

Sequences of 631,066 PM probes on the rice genome array downloaded from Affymetrix website were aligned against the pseudo-chromosomes of TIGR rice annotation version 5.0 using BLAT (Kent 2002) with parameters of minIdentity = 100, minMatch = 1 and stepSize = 5. Only probes with unique location to the rice genome were used to SFP confirmation and SNP analysis.

Rice and yeast SFP confirmation by SNP analysis in silico

Among 257,691 SNPs in OryzaSNP project (http://www.oryzasnp.org/), a total of 37,730 high-quality SNPs between Zhenshan 97 and Minhui63 were mapped to rice PM probes based on sequence identity and chromosomal location. The corresponding SNPs to SFP probes were identified as true detection positives. In addition, 100 Mb pseudo-sequences were downloaded from the same source and used to identify probes that are monomorphic between the two varieties.

To validate yeast SFPs, the whole genome shotgun sequences of *Saccharomyces cerevisiae* RM11-1a (RM) were downloaded from NCBI Nucleotide database by searching "txid285006[Organism:noexp]". Probe sequence alignments and polymorphisms were obtained by BLASTing (Altschul et al. 1990) the YG-S98 exemplar sequences from Affymetrix against the RM sequences using a custom perl script.

Calculation of minimum free energy of cRNA

The 75-bp reverse complementary sequences of BY and RM centered by each probe were extracted as cRNA sequences. UNAfold packages (Markham and Zuker 2005; Walter et al. 1994) were downloaded from author's website (http://dinamelt.bioinfo.rpi.edu/). The minimum RNA free energies of each extracted sequences were then calculated using 'hybrid-ss-min' command in UNAfold packages with parameters '-E -n RNA -t 45 -T 45' and the differences of free energies between the two genotypes were compared.

Results

SFP detection using median polish using rice data

Transcriptome profiling data of two indica rice cultivars Zhenshan 97 and Minghui 63 were collected from panicles at three developmental stages using the rice GeneChip genome array. The median polish method, adapted from Tukey (1977), was used in SFP detection. Thirty-six CEL files representing two genotypes, three developmental stages, and six replicates (three biological replicates, each with two technical replicates) were subjected to the median polish analysis. Only probe sets producing detectable hybridization signal, as indicated by "Present" calls by MAS 5.0 (Affymetrix Inc 2001), in at least one developmental stage of both rice cultivars were selected to construct data matrices. A total of 20,220 matrices each representing a positive probe set were constructed using signal intensities of perfect match (PM) probes after background correction and quantile normalization. probe set with more or less than 11 probes were removed from the analysis. The residuals from applying median polish were grouped by genotypes, and significant differential residual groups were identified using limma package (Smyth 2004) in the Bioconductor (Gentleman et al. 2004). The custom R scripts and microarray data can be downloaded in our website (see "Materials and methods").

Based on the assumption that the observed signal intensity of a PM probe can be decomposed to the corresponding transcript abundance level and a coefficient representing the binding affinity of the probe to the transcript target with a random error, the following model, proposed previously with demonstrated utility (Cui et al. 2005; Li and Hung Wong 2001), was adopted for our data analysis:

$$S_{gij} = I_{gi} + A_{gij} + \varepsilon_{gij} \quad (1)$$

where S_{gij} is the log-scaled observed signal intensity of the jth PM probe in the ith probe set hybridized to RNA derived from the gth genotype, I_{gi} represents the expression index of the transcript, A_{gij} the binding affinity coefficient of the probe, and ε_{gij} the random error. Different samples from the same genotype are considered to have the same A_{gij} but may differ in I_{gi}, whereas samples from different genotypes especially ones having sequence polymorphisms within the probe sequence would have different A_{gij} thus causing a SFP. For a $N \times M$ matrix composed of N samples and M probes within the ith probe set, median polish (Tukey 1977) decomposes each probe signal in the matrix to the following components:

$$S_{gij} = T_i + R_{gi} + C_j + E_{gij} \quad (2)$$

where T_i represents the overall value of the matrix, R_{gi} the effect of each row, C_j the effect of each column and E_{gij} the residuals. The relation of Eqs. 1 and 2 can be expressed as the following:

$$\begin{cases} \hat{I}_{gi} = T_i + R_{gi} \\ \hat{A}_{gij} = C_j + E_{gij} \end{cases} \quad (3)$$

That means I_{gi} can be estimated by the overall value T_i plus row effect of median polished R_{gi}. A_{gij} is directly proportional to the residual of median polished E_{gij}. For a probe hybridizing to two different genotypes, we expect to obtain two groups of residuals (E_{1ij}, E_{2ij}) due to different affinity effects (A_{1ij}, A_{2ij}). Therefore, we can deduce SFPs by applying median polish to the signal intensity matrix to get residuals and detecting the probes with differential residuals among different genotypes:

$$\Delta A_{ij} = \hat{A}_{2ij} - \hat{A}_{1ij} = E_{2ij} - E_{1ij} \quad (4)$$

Based on the above assumptions, we can extract the residual effects resulting from the different hybridization affinities of the two genotypes, irrespective of the expression levels and tissues.

In calculating the residuals for a $N \times M$ matrix composed of N samples (genotypes, tissues and replications) and M probes, the normalized intensity values in each row were subtracted by the median of the respective row, which eliminates the effects of the expression levels. The residuals in each column of the resulting matrix were then subtracted by the median of the column resulting in a new matrix, which removes the noise from different affinity of the probes. Such calculation was iterated until the sum of the absolute residuals in the matrix becomes stabilized (the sums of two successive iterations differed by <1%). SFPs between the varieties were identified based on the residuals of median polish using the limma package (Smyth 2004), with BH adjusted P values. An example of the effect obtained by median polish on SFP detection between the two rice cultivars is demonstrated in Fig. 1 using the probe set Os.10369.1.S1_a_at.

Using a threshold of BH adjusted $P \leq 0.001$, 6,655 SFPs from 3,131 probe sets, representing 15.5% of the total probe sets with "Present" calls, were identified between the two rice cultivars. Among them, 231 (7.4%) probe sets contained six or more SFPs.

SFP confirmation by SNP analysis

To evaluate the robustness of the method, detected SFPs were compared to the rice SNPs identified from oryzaSNP resequencing project (McNally et al. 2006). This project produced 100 Mb of unique and low copy rice genomic pseudo-sequences from 20 diverse rice lines, including Minghui 63 and Zhenshan 97, by microarray-based whole genome resequencing technology developed by Perlegen Sciences.

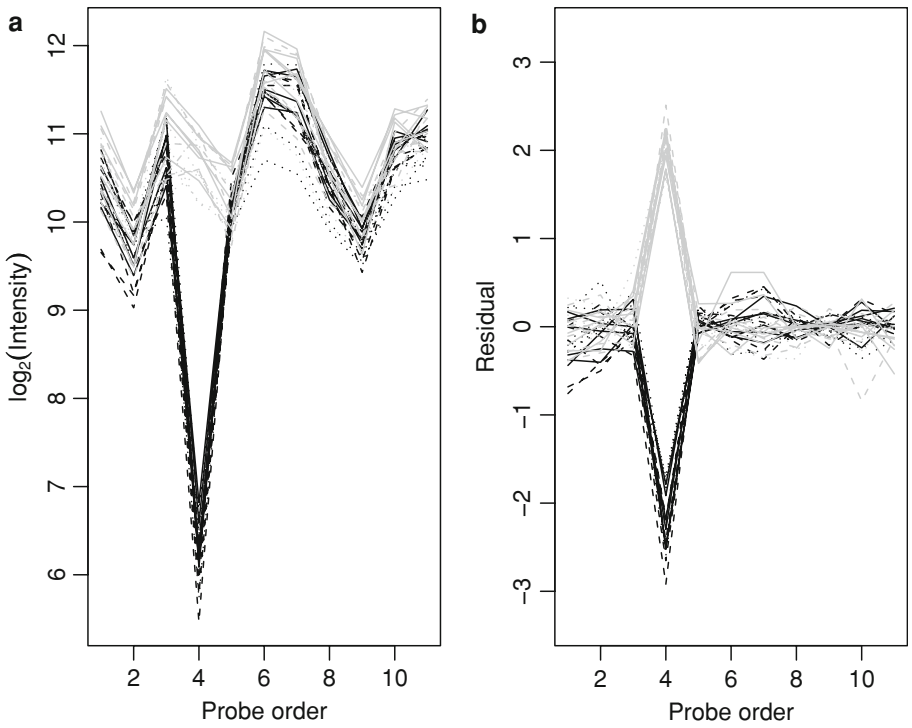

Fig. 1 Effectiveness of the median polish method in SFP detection from transcript profile data. **a** Normalized probe level signals of a probe set, Os.10369.1.S1_a_at. The *y* axis shows the background-corrected normalized $\log_2$ intensity and the *x* axis is the position of each probe of a probe set along the transcript. *Black lines* represent microarray results from Minghui 63 while *grey* are from Zhenshan 97. *Different line* types represent different tissues. **b** Same as **a** after median polish showing the effect on noise elimination of the probe signals and displaying the differential residuals of median polish at the SFP probe (the fourth probe) in the probe set

Based on the SNPs detected by OryzaSNP project, it was found that SNPs between these two varieties occurred in sequences corresponding to a total of 830 probe sets, 430 of which were resolved in our analysis, resulting in a false negative rate of 48.2% (Table 1). We checked the sequences for the false negative detections, and found that a SNP in the middle (the 13th base) of the 25-mer probe could be detected as a SFP with a false negative rate of 26.9%, while a SNP in the edge (the 1st and 25th bases) of the probe could be detected with a false negative rate of 90.2% (see Supplementary data Fig. 1 for the effect of SNP position for SFP detection). Among the 400 false negative SNPs, 191 (47.8%) were located within five bases from the ends (1st to 5th and 21st to 25th) of the probe sequences, while only 116 (29.0%) false negative SNPs were located within the 9 (9th to 17th) bases in the middle. Such positional effect of false negative detection is similar to the results from previous studies (Cui et al. 2005; Ronald et al. 2005; Rostoks et al. 2005).

Sequences for 569 of the 6,655 SFP detecting probes can be mapped to the oryzaSNP sequences, 430 (75.6%) of the mapped SFP probes have at least one SNP within the 25-base probe region between Minghui 63 and Zhenshan 97, and the remaining 139 (24.4%) SFPs did not show polymorphism between the two cultivars within the probe sequences (Table 1).

SFP detection in single tissue and multiple tissues

In order to assess the impact of multiple biological samples as replicates on SFP detection using median polish method, we predicted SFPs using data from rice samples collected at a single stage (single tissue), and compared the results to that from multiple tissues presented above. Different thresholds were used in different tissues to predict the same number of 6,655 SFPs as in the previous analysis using entire dataset. Totally, 91.4% SFPs detected by multiple tissues were detected by at least one single tissue. However, the single tissue analysis led to more false discovery and less sensitivity (Table 2), indicating positive gains of having multiple tissues in such analysis.

Comparison of the median polish method with other methods

For comparison, we re-analyzed our transcriptome data for SFP identification using the method of fitting linear model proposed by Rostoks et al. (2005). The exact method described previously was used, except SAM which needed large permutation (Tusher et al. 2001) was replaced with limma (Smyth 2004) to reduce computational demands of both time and memory. The detection power between limma and SAM using rice data was found to be highly similar: only 0.8% of the SFPs detected was different. In order to make the results comparable, detection threshold was adjusted to obtain a consistent number of 6,655 SFPs for both methods. By doing so, the cutoff probability was slightly different from 0.001 in this analysis. Totally, 4,937 (74.2%) of the 6,655 SFPs were the same between the two methods. Under the selected threshold, the median polish method had lower false positive, false negative and false

Table 1 Comparison of the detection power of different methods

Organism	Method	SFP number	False negative rate (%)	False positive rate (%)	False discovery rate (%)
Rice	Median polish[a]	6,655	48.2	0.90	24.4
	Fitting linear model	6,655	56.7	1.21	34.2
	NONE	6,655	58.7	1.28	36.6
Yeast	Median polish[a]	3,387	78.9	0.25	11.1
	Fitting linear model	3,387	81.4	0.48	21.2
	Median polish[b,c]	1,049	53.5	0.09	2.5
	PDNN[c]	1,049	55.3	0.22	6.2
Barley	Median polish[b]	10,504	23.7	3.23	18.8
	Fitting linear model	10,504	26.4	7.04	34.4

Method: median polish, a median polish step was employed to remove the transcript abundance effects from PM intensities; fitting linear model, obtaining the genotype effect at the probe level to identify SFPs by fitting linear model, Rostoks et al. proposed (Rostoks et al. 2005); PDNN, estimating the binding affinity of the probe to the transcript target based on the PDNN model, Ronald et al. proposed (Ronald et al. 2005); NONE, using PM intensities to identify SFPs directly

[a] BH adjusted *P* values less than 0.001 were used in median polish method for claiming SFPs, and the threshold for the counterpart of the comparison was adjusted to obtain an equal number of the predicted SFPs

[b] The threshold was adjusted to yield the same number of the predicted SFPs as in the published result (Ronald et al. 2005; Rostoks et al. 2005) for comparison

[c] The yeast SFPs detected by the PDNN method (and the corresponding median polish method) used only three replicates from the diploid genomes of the strains and a subset of the probe sets (which the authors regarded as robustly expressed probe sets)

Table 2 Efficiency of SFP detection between two rice cultivars using single tissue relative to all three tissues

Tissue[a]	ALL (%)	SBP (%)	PSP (%)	PCF (%)
False negative rate[b]	48.2	55.1	51.7	54.1
False positive rate[c]	0.90	1.18	1.05	1.08
False discovery rate[d]	24.4	32.9	28.9	30.6
Consistent SFP detection with ALL[e]	100	67.6	76.8	71.3

[a] ALL, all the three rice panicle tissues; SBP, PSP, and PCF, samples from three different developing stages of panicles (see "Materials and methods" for details). Each tissue was sampled with three biological replicates each with two technical repeats

[b] False negative: the rate in which SFPs were not predicted for probe sequences with known SNPs

[c] False positive: the rate of predicted SFPs among the sequences known not to contain SNPs

[d] False discovery: the rate of predicted SFPs whose probe sequences did not contain SNPs to the total number of SFPs. Both false positive and false discovery include ones with flanking SNPs

[e] The rate of consistence of SFPs detected with single tissues relative to multiple tissues

discovery rates compared to the liner model fitting method (Table 1). Moreover, even for the 139 SFPs called by median polish method that did not have SNPs in the probes, 85 (61.2%) were also called by the Rostoks' method, indicating that the results from these two analyses were highly consistent.

The SFPs predicted by the Rostoks' method were also validated by the same SNP data set from OryzaSNP project. Among the 6,655 SFP probes, 546 can be mapped to the oryzaSNP sequences, and 359 (65.8%) have SNPs within the 25-base probe regions, which is lower than that of the median polish method (75.6%). All but one of these SFP probes were detected by the median polish method. The remaining 187 (34.2%) were not supported by the oryzaSNP resequencing data. These results suggested that the median polish method has better detection sensitivity and accuracy than the method of fitting linear model.

We also directly used the original normalized PM intensities instead of residuals to identify SFPs. In doing so, all PM intensities of each probe were grouped by genotypes, and probes detecting significant difference between the two genotypes (SFPs) were identified using the limma package. When the detection threshold was also adjusted to obtain 6,655 SFPs, this method obviously resulted in lower sensitivity and accuracy than both the Rostoks' method and the median polish method (Table 1), indicating the positive gain of the median polish.

We also analyzed the barley microarray data composed of six tissues including radicle, root, leaf, embryo,

coleoptile and seedling crown, using the median polish method, and compared the results with the original analysis using the method of fitting linear model (Rostoks et al. 2005). It was shown that, with the same threshold of 10,504 SFPs and based on the available data, the median polish method is superior in both sensitivity and accuracy (Table 1).

We further compared these methods using yeast dataset from Ronald et al. (2005). Microarray data used for SFP detection between *S. cerevisiae* strains BY4716 (BY, an isogenic strain to the reference yeast) and RM11-1a (RM, a wild yeast strain) were downloaded from NCBI GEO (Barrett et al. 2007). The dataset included 12 microarrays, with three replicates from haploid and diploid genomes of each strain. Using the median polish method, 3,387 SFPs were detected between the two genotypes at $P < 0.001$ from all array data. As expected, for most (2,841) of the SFPs, BY showed higher binding ability than RM as estimated by average of residuals of median polish ($\bar{E}_{\mathrm{BY}} > \bar{E}_{\mathrm{RM}}$). Among 2,218 SFP detecting probes with available sequence information, SNPs were detected between the two strains within the 25 bases of 2,106 (95.0%) probes.

Interestingly, additional 546 SFPs were detected with unexpected lower residuals of median polish in BY than in RM ($\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$), indicating weaker binding ability in the reference strain BY than that of RM. We examined 186 SFP detecting probes from this group that have available sequence information, and found that only 32 (17.2%) of the 186 SFPs correspond to SNPs within the 25-base probe regions. However, the density distributions of the average absolute differences of residuals are similar between the SFP groups with $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$ residuals and with $\bar{E}_{\mathrm{RM}} < \bar{E}_{\mathrm{BY}}$ (Fig. 2), suggesting that there is an almost equal chance to detect SFPs with unexpected $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$.

When the same dataset was analyzed using the fitting linear model method with the same threshold of 3,387 SFPs, it produced higher false negative, false positive and false discovery rates than the median polish method (Table 1). A further comparison of the median polish method was also made against the method based on the positional-dependent nearest-neighbour (PDNN) model (Zhang et al. 2003) as in the original analysis, also demonstrating that the median polish method has better detection sensitivity and accuracy (Table 1).

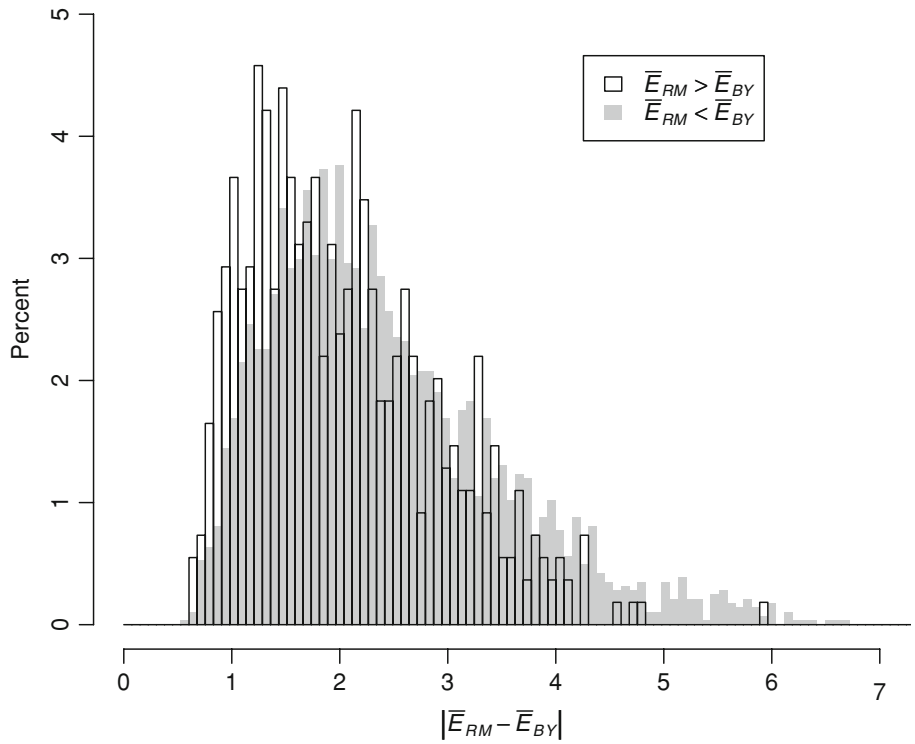

Fig. 2 The distribution of average absolute differences of residuals between two yeast strains BY and RM. Totally 3,387 SFPs between BY and RM were identified with BH adjusted $P < 0.001$. Their corresponding averages of residuals by median polish of BY ($\bar{E}_{\mathrm{BY}}$) and RM ($\bar{E}_{\mathrm{RM}}$) and their absolute differences ($|\bar{E}_{\mathrm{RM}} - \bar{E}_{\mathrm{BY}}|$) were calculated. The *x* axis shows $|\bar{E}_{\mathrm{RM}} - \bar{E}_{\mathrm{BY}}|$ and the *y* axis is the percentage of the distribution. The $|\bar{E}_{\mathrm{RM}} - \bar{E}_{\mathrm{BY}}|$ are grouped according to whether $\bar{E}_{\mathrm{RM}}$ is larger than $\bar{E}_{\mathrm{BY}}$ (*white* $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$) and (*grey* $\bar{E}_{\mathrm{RM}} < \bar{E}_{\mathrm{BY}}$). The residuals of median polish are considered as the affinity between transcript targets and the probe. The distribution of the average absolute differences in $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$ group are almost the same as in $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$ group, which means an equal chance to be detected as SFP although most SFPs in $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$ group are non-polymorphic probes

SFPs and probe flanking polymorphisms

As presented above, 24.4% of the SFPs, that could be repeatedly detected using different tissues and methods, did not have SNPs in the probe regions according to the OryzaSNP data. It was recently speculated that variations between different targets in flanking sequences adjacent to the probes may affect SFP detection (Borevitz et al. 2007; Zhu and Salmeron 2007). To assess in what degree probe flanking SNPs affect the detection of SFPs, we analyzed flanking SNPs of 5,088 probes with available flanking sequences longer than 20 bases at each end and the sequences corresponding to all these probes are monomorphic between the two rice varieties according to the data by the oryzaSNP resequencing project (McNally et al. 2006). SNPs were found within flanking 20 bases between Zhenshan 97 and Minghui 63 for 217 probes. Among them, 77 probes have flanking SNPs in the immediate adjacent ten bases, and 142 probes have SNPs within the flanking 11th to 20th base region. Two probes with SNPs in both immediate adjacent 10 bases and 11th to 20th bases were removed from the analysis.

It was found that the probe flanking SNPs were significantly associated with the SFP calls (Table 3; Fig. 3a). Among the SFPs detected at $P < 0.001$, SNPs occurred with a frequency of 14.3% in the flanking ten bases, a 9.7-fold enrichment comparing to the background frequency (1.5%) resulting in $P = 1.40\text{e-}04$ by Fisher's exact test (FET). SFP calls were further enriched to 17-fold when the threshold stringency increase to $P < 0.00001$ (4/16, FET, $P = 6.94\text{e-}05$). About a twofold enrichment (5.7%) of SNPs was also observed in the flanking 11th to 20th base regions among the SFP calls at $P < 0.001$ compared with the background (2.8%).

We again used yeast microarray data (Ronald et al. 2005) to validate the above observation (Fig. 3b). We particularly focused on the erratic SFP probes with negative residuals ($\bar{E}_{RM} > \bar{E}_{BY}$) detected by reference yeast strain BY. A total of 4,273 probes without polymorphisms within the 25 bases but with SNPs in the flanking 20 bases of the probes were selected for the analysis (Table 3). The concurrence of SFPs with probe flanking SNPs was found to be significantly higher than in the background. There are only 4.6% (2,127) of the probes having SNPs in flanking ten bases among 46,595 monomorphic probes, compared to 15.1% (13) among the 86 significant SFPs detected at $P < 0.001$ (FET, $P = 1.39\text{e-}04$). However, the rate of SNPs was not any higher among the SFPs than the background in flanking 11th to 20th base regions (1/86, FET, $P = 0.98$). These results indicated that probe flanking polymorphisms were partially responsible for SFPs, including SFPs resulting from negative residuals.

SFPs and nucleotide composition of probe flanking SNPs

We also investigated the possible effect of the nucleotide compositions of the flanking SNPs on hybridization signals of the SFPs. We first categorized the yeast flanking SNPs into two groups according to their relative residuals of median polish in RM and BY: $\bar{E}_{RM} > \bar{E}_{BY}$ or $\bar{E}_{RM} < \bar{E}_{BY}$. We then calculated the percentage of the four nucleotides in yeast RM11-1a genotype, and plotted these percentages along the different thresholds of BH adjusted P value generated from SFPs detection in each group (Fig. 4). We found that adenines appear as predominant form of purines in the $\bar{E}_{RM} > \bar{E}_{BY}$ group under the same or more stringent P values while guanines display opposite effects, suggesting possible influence of adenines and guanines in the flanking SNPs to the probe binding affinity. However, compared to the background, the observed differences are relatively small although they are statistically significant (when using subset flanking SNPs under SFPs detection $P < 0.05$, $\bar{E}_{RM} < \bar{E}_{BY}$: $\chi^2 = 7.00$, $df = 3$, $P = 0.072$; $\bar{E}_{RM} > \bar{E}_{BY}$: $\chi^2 = 6.88$, $df = 3$, $P = 0.076$). Compared with purines, the curves of pyrimidines (thymines and cytosines) showed weaker correlation between the SNP nucleotide composition and the flanking SFP formation.

RNA secondary structure induced by flanking polymorphisms

It was speculated that the secondary structure of RNA molecules in a solution may interfere with the binding to

Table 3 SFPs caused by flanking SNPs revealed with different thresholds

Perfect matched probe	Rice			Yeast		
		$P < 0.05$[a]	$P < 0.001$[a]		$P < 0.05$[a]	$P < 0.001$[a]
Total	5,086[b]	131	35	46,595[b]	819	86
With SNPs in flanking 0–10 bp	75 (1.5%)	9 (6.9%)	5 (14.3%)	2,127 (4.6%)	74 (9.0%)	13 (15.1%)
With SNPs in flanking 11–20 bp	140 (2.8%)	10 (7.6%)	2 (5.7%)	2,146 (4.6%)	44 (5.4%)	1 (1.2%)

[a] Number of SFPs detected at the BH adjusted P values of 0.05 and 0.001, respectively

[b] Number of probes that are monomorphic in the probe region with flanking sequences available in public database. For the yeast data, we only selected the probes with $\bar{E}_{RM} > \bar{E}_{BY}$

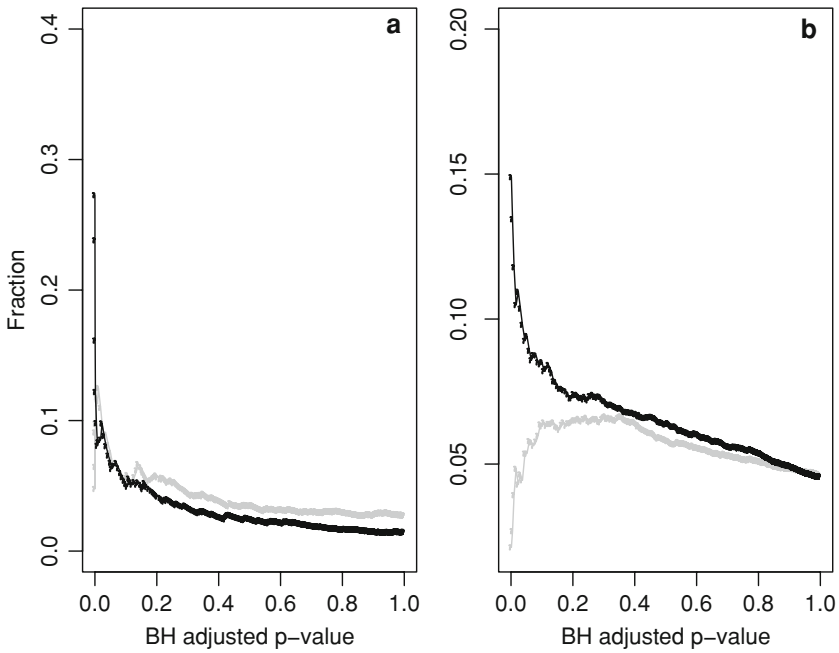

Fig. 3 The fraction distribution of non-polymorphic probes with flanking SNPs resolved with different statistical thresholds in **a** rice and **b** yeast. The fractions of probes with flanking SNPs were calculated using different thresholds of BH-adjusted *P* value generated from SFP detection. The *x* axis shows the BH-adjusted *P* values and the *y* axis is the fraction. The surveys in the flanking 1–10 base region and 11–20 base region are colored in *black* and *grey*, respectively

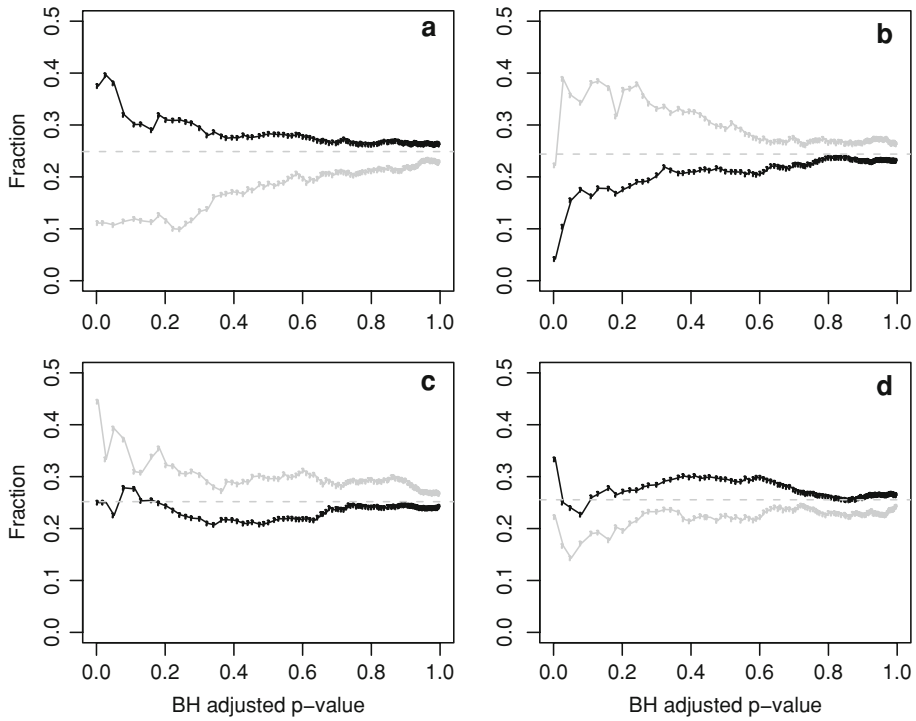

Fig. 4 The fraction distribution of nucleotide content of SNPs flanking non-polymorphic probes resolved with different statistical thresholds. Non-polymorphic probes with unique SNPs in the flanking 1–10 bases were selected and the nucleotide content of these SNPs of yeast RM11-1a genotype were then surveyed using different thresholds of BH-adjusted *P* values generated from SFP detection. The *x* axis shows the BH adjusted *P* values and the *y* axis is the fraction of certain nucleotide base. SNPs are grouped according to whether $\bar{E}_{\mathrm{RM}}$ is larger than $\bar{E}_{\mathrm{BY}}$ (*black* $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$) and (*grey* $\bar{E}_{\mathrm{RM}} < \bar{E}_{\mathrm{BY}}$). **a** Adenines appear as predominant form of purines in the $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$ group under more stringent BH-adjusted *P* values, while **b** guanines show the opposite trend, suggesting possible influence of adenines and guanines in the flanking SNPs to the probe binding affinity. Cytosines **c** and thymines **d** show weaker correlation between the SNP nucleotide composition and the flanking SFP formation

probes (Southern et al. 1999). To examine the changes of cRNA secondary structure induced by flanking polymorphisms, we used the UNAfold software (Markham and Zuker 2005; Walter et al. 1994) to calculate the minimum free energies (ΔG) of 75-bp cRNA sequences that flank the center of each probe extracted from BY and RM sequences. A total of 3,439 polymorphic flanking sequence pairs with monomorphic probe region and polymorphic flanking ten bases were obtained. We also divided these sequence pairs into two groups according to their relative residuals of median polish: $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$ or $\bar{E}_{\mathrm{RM}} < \bar{E}_{\mathrm{BY}}$. We then calculated the proportion of sequence pairs in which the RNA free energy of RM sequences were greater than BY sequences (proportion of $\Delta G_{\mathrm{RM\text{-}BY}} > 0$) according to different SFP detection P values. The results showed that there was no change of $\Delta G_{\mathrm{RM\text{-}BY}}$ for 738 (21.5%) of the sequences. A threshold of 50% quantile (absolute value of $\Delta G_{\mathrm{RM-BY}} > 0.60$ kJ) was applied to filter sequence pairs with different minimum RNA free energy between BY and RM (see the Supplementary Fig. 2 for the distribution of $\Delta G_{\mathrm{RM-BY}}$). Although it is difficult to confirm all of the predicted changes in RNA secondary structures, as the accuracy of predictions of RNA secondary structures is only 74% (Walter et al. 1994), we found that the minimum free energy of RNA is correlated positively to the binding affinity (Fig 5). Using a threshold of BH adjusted P values less than 0.05, the percentage of sequence pairs with $\Delta G_{\mathrm{RM-BY}} > 0$ (72.5%) in the $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$ group is significantly higher than the background level (54.1%, FET, $P = 0.012$). Reversely, only 28.6% of the sequence pairs with $\Delta G_{\mathrm{RM-BY}} > 0$ were found in the $\bar{E}_{\mathrm{RM}} < \bar{E}_{\mathrm{BY}}$ group. However, the difference is statistically insignificant compare to the background level (46.5%, FET, $P = 0.29$) due to the less frequent occurrence and subsequently smaller sample size ($n = 7$, at SFPs detection $P < 0.05$). Unfortunately, we are unable to conduct similar analysis in rice, as the flanking sequences are not available presently.

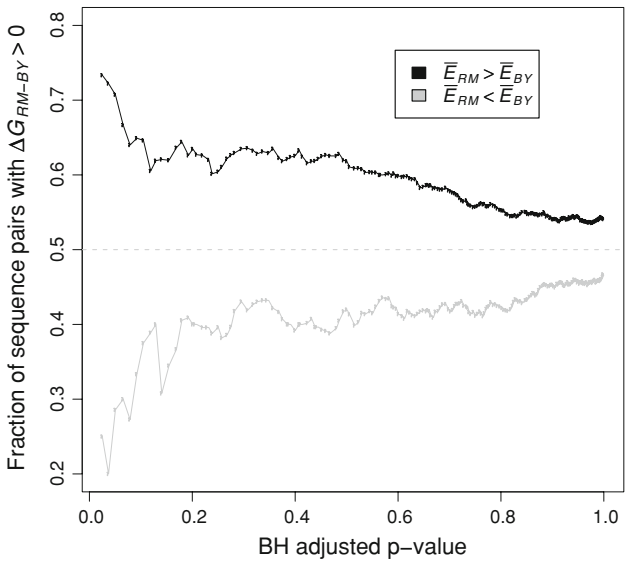

Fig. 5 The fraction distribution of the free energy changes in non-polymorphic probes with flanking SNPs resolved with different statistical thresholds. Non-polymorphic probes with SNPs in the flanking 1–10 bases were selected and their 75-bp cRNA sequence pairs flanking the center of these probes were extracted from yeast BY and RM sequences. The minimum RNA free energies of the sequence pair (ΔG_{BY}, ΔG_{RM}) are then calculated and the difference between ΔG_{BY} and ΔG_{RM} are calculated by subtracting ΔG_{BY} from ΔG_{RM} (denoted as $\Delta G_{\mathrm{RM\text{-}BY}}$). The fractions of $\Delta G_{\mathrm{RM-BY}} > 0$ were then surveyed under different thresholds of BH-adjusted P values generated from SFP detection. The x axis shows the BH-adjusted P values and the y axis is the fraction of $\Delta G_{\mathrm{RM-BY}} > 0$ with the statistical thresholds. Sequence pairs are divided into two groups according to whether $\bar{E}_{\mathrm{RM}}$ is larger than $\bar{E}_{\mathrm{BY}}$ (*black* $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$) and (*grey* $\bar{E}_{\mathrm{RM}} < \bar{E}_{\mathrm{BY}}$). There is a trend that the proportion of sequence pairs with $\Delta G_{\mathrm{RM-BY}} > 0$ increases in the group of $\bar{E}_{\mathrm{RM}} > \bar{E}_{\mathrm{BY}}$ while the proportion decreases in the group of $\bar{E}_{\mathrm{RM}} < \bar{E}_{\mathrm{BY}}$ with more stringent thresholds

Discussion

Median polish as a high performing method for SFP detection

Molecular interactions between probes and targets are the basis for microarray detection. Both binding affinity between targets and probes and target abundance contribute to the output signals measured by microarray probes. Ideally, one of the two factors ought to be fixed or ignored in order to measure the other. For example, in transcriptome analysis, only the abundance of the transcript targets is considered as the variable. In the sequence analysis, in contrast, the sequence differences between targets and probes are considered as the major variable while the abundance of the targets is assumed similar. However, such a highly simplified model is not applicable in a variety of microarray applications. SFPs are consequence of sequence difference between corresponding targets from two samples detected by the same probes, providing opportunity to detect sequence polymorphism, as well as challenge to ensure accuracy in transcript abundance measurement in a transcript analysis.

SFP detection from RNA derivative targets by microarray has the advantage of measuring both transcript abundance and sequence variation at the same time (Cui et al. 2005; Luo et al. 2007; Ronald et al. 2005; Rostoks et al. 2005; West et al. 2007). It makes the SFP detection more cost effective and possible in complex genomes by using expressed transcripts as genome complexity reduction method (Cui et al. 2005; Gore et al. 2007; Rostoks et al. 2005; Zhu and Salmeron 2007). However, the algorithms used for analyzing such data requires not only

identifying differential hybridization signals but also dissecting contributing factors to these differential signals to ensure sufficient detection power and accuracy.

We adapted a median polish based method to identify SFPs using expression data from multiple tissues. Our study suggested that the median polish method has superior performance in SFP detection as evidenced by validation using data from multiple genomes with different composition, structure, and complexity (Table 1). An additional advantage of this method is that most methods for SFP prediction need specific design of microarray experiments (generally two groups under the same experimental conditions with several repeats in each group) (Cui et al. 2005; Das et al. 2008; Luo et al. 2007; West et al. 2006), our median polish based method provides some uniquely useful features for SFP detection. This method works for dissecting the expression data into and eliminate the effects of transcript abundance and variations of binding affinity, thus perturbation of transcript abundance resulting from different experimental conditions and data types had little interference with the sensitivity and accuracy of the SFP detection. This method can also significantly reduce the computing time required to perform the same SFP analysis using the same data set, thus enable one to perform a SFP analysis using more complex data set. The computing time required for analyzing the rice data set in this study is less than 1 min by the median polish method and 21 min by the fitting linear model. The replacement of SAM with limma could further reduce the memory requirements and computing time by 50 times.

Complexity of SFP formation and the contribution from probe flanking SNP

There are many factors influencing hybridization signal from a RNA target, including whether the RNA target was expressed in the sample, the abundance of the RNA target, the GC content of the binding region, and the presence of mismatch bases between the target and probe and their location (Rostoks et al. 2005; Zhu and Salmeron 2007). It was generally considered that mismatch bases within probe sequences lead to reduced hybridization intensities and formation of SFPs in comparative studies (Fig. 6).

However, the interaction between short oligonucleotide probe and target is complex (Naef et al. 2002b) and affected by many variables in addition to the above described sequence specificity and abundance of targets and probes. Such complexity can be at least demonstrated by the following exceptional phenomena. First, probes could produce greater hybridization signals from targets with less complementary sequences. It is well known that a significant fraction of MM probes, which are designed for probe non-specific hybridization, can bind targets better

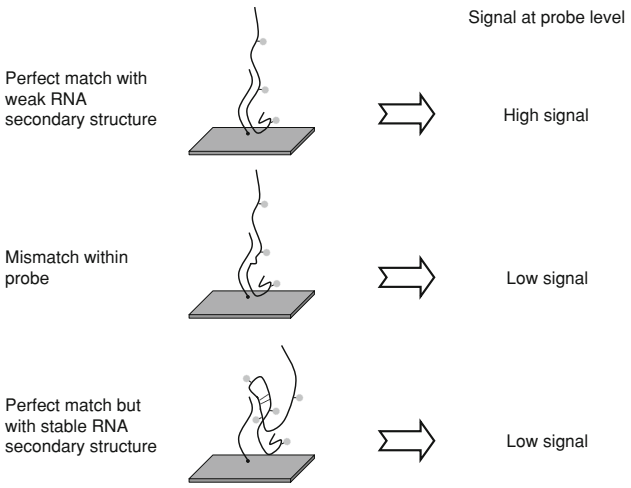

Fig. 6 Probe–target interaction and contribution to hybridization signals. When target abundance is presumed to be at similar level, target that matches probe perfectly and with weak secondary structure produces a relative high signal, while target with mismatch within probe region produces a relative low signal. When the flanking polymorphisms can make more stable RNA internal secondary structure, the binding affinity between probes and targets may lead to reduced signal

than the PM probes (Naef et al. 2002a, b; Zhang et al. 2003). In this study, 32 yeast SFPs with confirmed mismatch bases between RM targets and probes was found to have greater residuals of median polish in RM and more than half of them (18/32) was found to have greater hybridization signal in RM. Second, probes could produce different hybridization signals from targets with identical perfect complementary sequences. A total of 266 probes detected statistically significant differential signals from targets of RM and BY strains even these targets have sequences perfectly matched the probes.

The interference of biotinated fluorescent labels (Naef and Magnasco 2003) and stacking free energies of unlabeled nucleic acids in solution (Carlon et al. 2006) are among the hypothesized contributing factors to the first phenomenon. However, neither of the hypotheses was fully supported by experimental evidence (Wang et al. 2007).

The second phenomenon may relate to the influence of the mismatch bases in the flanking region (Fig. 6). Mismatch bases in the flanking region of a probe are generally considered to have less direct effect on SFP detection and thus less studied. However, our results from analyzing both rice and yeast data clearly demonstrated that the probe flanking SNPs could significantly contribute to SFP formation as over 15% of non-polymorphic SFPs are associated with flanking SNPs, of which the background contribution was less than 5%.

To uncover the molecular basis of the SFP formation by the flanking SNPs, we surveyed the nucleotide composition of the flanking SNPs and confirmed that adenines are over-

represented and guanines are underrepresented in the flanking SNPs that lead to higher binding affinity. We also calculated minimum free energies of 75-bp cRNA sequences flanking the center of each probe and compared their difference between targets. We found that the hybridization affinity was correlated positively to the minimum free energy of cRNA. The changes of predicted minimum free energy of cRNA are significant. These observations suggest that the flanking SNPs could lead to the formation of SFPs by changing minimum free energies of unlabeled RNA targets: a lower minimum free energy may lead to more stable RNA internal secondary structure thus reducing the hybridization affinity between probes and targets.

To further investigate possible causes for the SFPs, we inspected 28 rice SFPs (Table 3) that were detected at $P < 0.001$ but had no SNPs in either the probes or the flanking 20 bp regions, 24 of which were genes with EST/cDNA support in the TIGR database. Ten (41.7%) of the 24 genes have more than one gene model presumably due to alternative splicing, much higher than the estimated 21.2% alternative splicing level of the rice genome (Wang and Brendel 2006), suggesting the likelihood that alternative splicing may also contribute to the SFP detection. Moreover, 7 of the 28 SFP detecting probes were the last ones (No. 11) in the respective probe sets. Recently, unexpected 3′ alternative termination of gene transcription was observed by RNA sequencing of mammal and yeast genomes (Carninci et al. 2005; Nagalakshmi et al. 2008). Considering that the microarray we used was Affymetrix 3′ expression arrays, in which the probes primarily target the 3′ end of the genes, these observations might also suggest possible differences in 3′ alternative termination between the two genotypes. Furthermore, the widely reported copy number variations (insertions/deletions) of small DNA segments exist not only in human and rat (Guryev et al. 2008; Stranger et al. 2007), but also in plants and other taxa (Clark et al. 2007; Gresham et al. 2006). These copy number variations along with alternative splicing and/or 3′ alternative termination may affect the probe signal intensities, thus might have been detected as SFPs in this analysis while not necessarily detected by sequencing, thus contributing to the discrepancies between the sequence data and SFP detection.

Usefulness of the detected rice SFPs

Rice is a staple cereal for a large segment of the world population. The two varieties used in this study, Zhenshan 97 and Minghui 63, are the parents of Shanyou 63, the most widely grown hybrid in China. Hundreds of QTLs for a large array of traits and yield heterosis have been mapped using populations derived from this cross. Recently, a gene underlying a major QTL for pleiotropic effects on heading date and yield potential was cloned (Xue et al. 2008). Although this is probably the most intensively studied cross in rice genetic mapping, several large gaps (>20 cM) still exist in the genetic map due to lack of polymorphic markers in these regions (Hua et al. 2002; Xing et al. 2002), despite large efforts in adding more markers. Such large gaps would have certainly affected the completeness of the information for genetic mapping. The discovery of 6,655 SFPs in this study representing 3,131 rice unique genes with a high validation rate supported by large scale resequencing data, could generate an extra-high density genetic map, thus greatly facilitate gene discovery in this population.

In conclusion, median polish has superior performance in SFP detection. This method avoids constructing complex linear models thus could be used to genotype recombinant inbred lines or double haploid population which cannot be fit into linear models easily. Furthermore, it could be used to remove the influence of SFPs in the expression analyses using GeneChip arrays to improve accuracy of measurement. In this aspect, it would be especially powerful when the method is combined in the same automated workflows with the robust multi-array average (RMA) (Irizarry et al. 2003), one of the most adapted algorithm for expression analysis using GeneChip microarrays, which uses the same algorithm.

We confirmed that probe flanking SNPs could lead to SFPs in yeast and rice genomes that significantly different in genome size, composition, and complexity level. This finding has impact not only in SFP detection and its application in genetic analysis, but also in expression analysis using GeneChip and other short oligonucleotide probe arrays. These probe flanking SNPs could affect the binding affinity by changing the minimum free energy of the unlabeled RNA target molecules. Our results also indicated the complex compositions of SFP formation, for a large of SFPs still cannot be explained by either probe SNPs or flanking SNPs.

To the best of our knowledge, our study was the first application of detecting SFPs using expression data in rice. The success of our study demonstrated the applicability of this approach. The detected rice SFPs would greatly facilitate the effort of saturating the genome with molecular markers.

Acknowledgments We thank Dr. James Ronald and Dr. Rachel B. Brem for help and suggestions in yeast data. This work was supported by grants from the National Special Key Project of China on Functional Genomics of Major Plants and Animals, and the National Natural Science Foundation of China.

References

Affymetrix Inc (2001) GeneChip expression analysis technical manual

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D et al (2007) NCBI GEO: mining tens of millions of expression profiles-database and tools update. Nucleic Acids Res 35:D760–D765

Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T et al (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res 13:513–523

Borevitz JO, Hazen SP, Michael TP, Morris GP, Baxter IR et al (2007) Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 104:12057–12062

Carlon E, Heim T, Wolterink JK, Barkema GT (2006) Comment on "Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays". Phys Rev E 73:063901

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC et al (2005) The transcriptional landscape of the mammalian genome. Science 309:1559–1563

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G et al (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science 317:338–342

Cui X, Xu J, Asghar R, Condamine P, Svensson JT et al (2005) Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit. Bioinformatics 21:3852–3858

Das S, Bhat PR, Sudhakar C, Ehlers JD, Wanamaker 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

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy-analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20:307–315

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80

Gore M, Bradbury P, Hogers R, Kirst M, Verstege E et al (2007) Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. Crop Sci 47:S-135–S-148

Gresham D, Ruderfer DM, Pratt SC, Schacherer J, Dunham MJ et al (2006) Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. Science 311:1932–1936

Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA et al (2008) Distribution and functional impact of DNA copy number variation in the rat. Nat Genet 40:538–545

Hua JP, Xing YZ, Xu CG, Sun XL, Yu SB, Zhang Q (2002) Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. Genetics 162:1885–1895

Huang Y, Zhang L, Zhang J, Yuan D, Xu C et al (2006) Heterosis and polymorphisms of gene expression in an elite rice hybrid as revealed by a microarray analysis of 9198 unique ESTs. Plant Mol Biol 62:579–591

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664

Kumar R, Qiu J, Joshi T, Valliyodan B, Xu D, Nguyen HT (2007) Single feature polymorphism discovery in rice. PLoS ONE 2:e284

Li C, Hung Wong W (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol 2:RESEARCH0032

Luo ZW, Potokina E, Druka A, Wise R, Waugh R, Kearsey MJ (2007) SFP genotyping from affymetrix arrays is robust but largely detects *cis*-acting expression regulators. Genetics 176:789–800

Markham NR, Zuker M (2005) DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res 33:W577–W581

McNally KL, Bruskiewich R, Mackill D, Buell CR, Leach JE, Leung H (2006) Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. Plant Physiol 141:26–31

Naef F, Hacker CR, Patil N, Magnasco M (2002a) Characterization of the expression ratio noise structure in high-density oligonucleotide arrays. Genome Biol 3:PREPRINT0001

Naef F, Lim DA, Patil N, Magnasco M (2002b) DNA hybridization to mismatched templates: a chip study. Phys Rev E 65:040902

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

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D et al (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349

Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M (2008) Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J 53:90–101

Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L (2005) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. Genome Res 15:284–291

Rostoks N, Borevitz JO, Hedley PE, Russell J, Mudie S et al (2005) Single-feature polymorphism discovery in the barley transcriptome. Genome Biol 6:R54

Seheult AH, Tukey JW (2001) Towards robust analysis of variance. Data Analysis from Statistical Foundations. Nova Publishers, New York, pp 217–244

Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article3

Southern E, Mir K, Shchepinov M (1999) Molecular interactions on microarrays. Nat Genet 21:5–9

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C et al (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853

Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Menlo Park

Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98:5116–5121

Walter AE, Turner DH, Kim J, Lyttle MH, Muller P et al (1994) Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. Proc Natl Acad Sci USA 91:9218–9222

Wang BB, Brendel V (2006) Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci USA 103:7175–7180

Wang Y, Miao ZH, Pommier Y, Kawasaki ES, Player A (2007) Characterization of mismatch and high-signal intensity probes

associated with Affymetrix genechips. Bioinformatics 23:2088–2095

West MA, van Leeuwen H, Kozik A, Kliebenstein DJ, Doerge RW et al (2006) High-density haplotyping with microarray-based expression and single feature polymorphism markers in *Arabidopsis*. Genome Res 16:787–795

West MA, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW et al (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. Genetics 175:1441–1450

Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S et al (1998) Direct allelic variation scanning of the yeast genome. Science 281:1194–1197

Xing Z, Tan F, Hua P, Sun L, Xu G, Zhang Q (2002) Characterization of the main effects, epistatic effects and their environmental interactions of QTLs on the genetic basis of yield traits in rice. Theor Appl Genet 105:248–257

Xue W, Xing Y, Weng X, Zhao Y, Tang W et al (2008) Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. Nat Genet 40:761–767

Zhang L, Miles MF, Aldape KD (2003) A model of molecular interactions on short oligonucleotide microarrays. Nat Biotechnol 21:818–821

Zhu T, Salmeron J (2007) High-definition genome profiling for genetic marker discovery. Trends Plant Sci 12:196–202