Plant Biotechnology Journal (2006) **4**, pp. 489–498 doi: 10.1111/j.1467-7652.2006.00198.x

Genetic diversity contribution to errors in short oligonucleotide microarray analysis

Matias Kirst[1,*,†], Rico Caldo[2], Paula Casati[3,‡], Gene Tanimoto[4], Virginia Walbot[3], Roger P. Wise[2,5] and Edward S. Buckler[1,6]

[1]*Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853-2703, USA*
[2]*Department of Plant Pathology, Iowa State University, Ames, IA 50011, USA*
[3]*Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA*
[4]*Affymetrix, Inc., Santa Clara, CA 95051, USA*
[5]*USDA-ARS, Corn Insects and Crop Genetics Research, Iowa State University, Ames, IA 50011, USA*
[6]*USDA-ARS, Plant, Soil, and Nutrition Research Unit, Cornell University, Ithaca, NY 14853, USA*

Received 26 January 2006;
revised 17 April 2006;
accepted 20 April 2006
**Correspondence* (fax +1 352 846 1277; e-mail: mkirst@ufl.edu)
†*Present address*: School of Forest Resources and Conservation, University of Florida, PO Box 110410, Gainesville, FL 32611, USA
‡*Present address*: Centro de Estudios Fotosinteticos y Bioquimicos (CEFOBI), Suipacha 531, 2000 Rosario, Argentina

Keywords: Affymetrix, genetic diversity, maize, microarray, type I error.

Summary

DNA arrays based on short oligonucleotide (≤ 25-mer) probes are being developed for many species, and are being applied to quantify transcript abundance variation in species with high genetic diversity. To define the parameters necessary to design short oligo arrays for maize (*Zea mays* L.), a species with particularly high nucleotide (single nucleotide polymorphism, SNP) and insertion-deletion (indel) polymorphism frequencies, we analysed gene expression estimates generated for four maize inbred lines using a custom Affymetrix DNA array, and identified biases associated with high levels of polymorphism between lines. Statistically significant interactions between probes and maize inbreds were detected, affecting five or more probes (out of 30 probes per transcript) in the majority of cases. SNPs and indels were identified by re-sequencing; they are the primary source of probe-by-line interactions, affecting probeset level estimates and reducing the power of detecting transcript level variation between maize inbreds. This analysis identified 36 196 probes in 5118 probesets containing markers that may be used for genotyping in natural and segregating populations for association gene analysis and genetic mapping.

Introduction

High-throughput DNA arrays produced by *in situ* synthesis of short (≤ 25-mer) oligonucleotide sequences (Affymetrix, Santa Clara, CA, USA) are one of the most popular platforms for massive parallel gene expression analysis. Such microarrays are now available for human, many plant and animal model species, and an increasing number of agricultural crops. Although the utility of such DNA arrays in allied species has been shown to be limited (Hsieh *et al*., 2003; Gilad *et al*., 2005), their suitability for wide application within a species has typically not been questioned. Nonetheless, several commercially important crop plants are intraspecifically far more diverse than the most broadly studied mammalian species. Maize, the most valuable crop in the USA, is one of the most genetically diverse cultivated plants and shows both single nucleotide polymorphism (SNP) and insertion-deletion (indel) polymorphism frequencies much higher than most eukaryotes (Tenaillon *et al*., 2001). Therefore, the question arises as to whether this high level of intraspecific genetic diversity will prevent or reduce the confidence in gene expression analysis using oligonucleotide-based DNA arrays.

To define inconsistent probes, Affymetrix has typically designed a set of several 25-mer probes (typically 11) matching each gene [perfect match (PM) probes], accompanied by a second set of probes [the mismatch (MM) set] with a single substitution at the 13th base. The purpose of the MM probe is to approximate background for the subtraction of

non-specific and cross-hybridization effects. Probes that report unexpected results (outliers) can originate from cross-hybridization to non-target genes, contaminated regions in the DNA array and other artefacts (Li and Wong, 2001). Signal estimation methods, such as Robust Multi-Array Analysis (RMA) and Probe Logarithmic Intensity Error Estimation (PLIER), are typically used to summarize probe level results while accounting for outlier probes. Notably, DNA sequence variation can limit the value of certain probes when polymorphisms exist between the biological sample and the reference genotype used in the DNA array design (Ronald *et al.*, 2005; Rostoks *et al.*, 2005). For species with low genetic diversity, such polymorphisms may be negligible and may affect only a small number of probes (e.g. human $\theta_{total} < 0.001$; Cargill *et al.*, 1999; Halushka *et al.*, 1999). For highly diverse plant crops, however, DNA sequence variation can be significant. In maize, two inbred lines and landraces vary on average in 1/139 bases in coding regions ($\theta_{coding} = 0.0072$). When non-coding regions are included, Tenaillon and colleagues found one out of every 28 bases to be polymorphic in a sample of 25 maize inbred lines and landraces (Tenaillon *et al.*, 2001). This polymorphism estimate did not account for small insertions and deletions, which are found in most sequences. Therefore, genetic polymorphism in highly diverse species could substantially impact the level of detection in a significant proportion of probes, particularly as the populations sampled are more distantly related to the standard used in probe design. Nearly all of the expressed sequence tags (ESTs) and genome sequence tags for maize are drawn from a handful of inbred lines, principally B73, W23 and Oh43A; these sequence data are the basis for the design of probesets. On average, more than 50% of 25-mer probes would be expected to contain a locus that is polymorphic in an experiment including 25 maize lines. Even in a smaller sample size of four lines, such as used in this study, 30% of probes are likely to contain a polymorphic locus.

In this study, we contrasted gene expression between four maize lines using a short oligonucleotide microarray comprising 8346 genes, each represented by thirty 25-mer probes. Analysis of the interactions between individual maize lines and each probe identified substantial bias in the identification of differentially expressed genes. This could be attributed to interspecific genetic diversity. Although polymorphism in maize can involve the acquisition of statistically robust gene expression data for diverse lines, the sensitivity of the short oligonucleotide probes is shown to yield a potentially fast, reliable and cost-effective approach for SNP genotyping of highly diverse maize lines, and is predicted to demonstrate similar utility with other species.

Results

Biases in gene expression analysis of maize

To evaluate the role of genetic diversity in maize transcript analysis, we conducted an experiment to characterize transcript abundance in four US maize lines: B73, Mo17, Wf9-BG (T-cytoplasm) and W23. Leaf RNA was labelled and hybridized to a custom Affymetrix GeneChip® maize array (Affymetrix) comprising 8346 genes, each represented by a probeset of 30 PM and 30 MM probes. Approximately half of the probesets were designed against sequences from the maize line B73, and the remaining originated from all the maize sequences available in GenBank and dbEST (referred to hereafter as 'B73 probesets' and 'Maize pool probesets', respectively). The B73 probesets allowed us to compare the performance of probes based on one reference genome (B73) in the other three maize inbred lines. B73 has been classified as derived from the stiff stalk group, whereas Mo17 and Wf9-BG are non-stiff stalks.

For each probeset, the signal estimate computed on all 30 PM probes (including three biological replications with two technical replications each) was used to identify significant differences in relative transcript abundance between the four maize lines, after data analysis using the Affymetrix algorithms MAS5 (Microarray Suite 5.0) and PLIER. A one-way analysis of variance (ANOVA) was used to estimate individual line effects on measured signal estimates. Data distribution showed bias towards higher signal estimates in the genotype used as the reference genome for the construction of the DNA array (Figure 1A). Consequently, an erroneous conclusion might be made of higher relative expression in the reference genotype. In particular, for B73, signal estimates were systematically lower in the Maize pool probesets and higher in the B73 probesets, even after normalization and summarization using MAS5 and PLIER. Next, we identified the probesets which had signal estimates significantly higher in one line relative to the other three (Figure 1B). A similar proportion of probesets was identified as having higher signal estimates in each of the lines Mo17, W23 and Wf9-BG (~28%) in the Maize pool probeset. The line B73 represented a lower proportion of these (14%). An opposite trend was observed in the B73 probeset, where the signal estimate was significantly higher in B73 for 53% of the probesets. A naive interpretation of the results from the B73 probeset would be that the number of genes significantly more highly expressed in B73 was 3.4 times higher than the average observed in the other line – 157 genes in B73 compared with an average of 47 in the other lines. This interpretation is a statement that properties of the

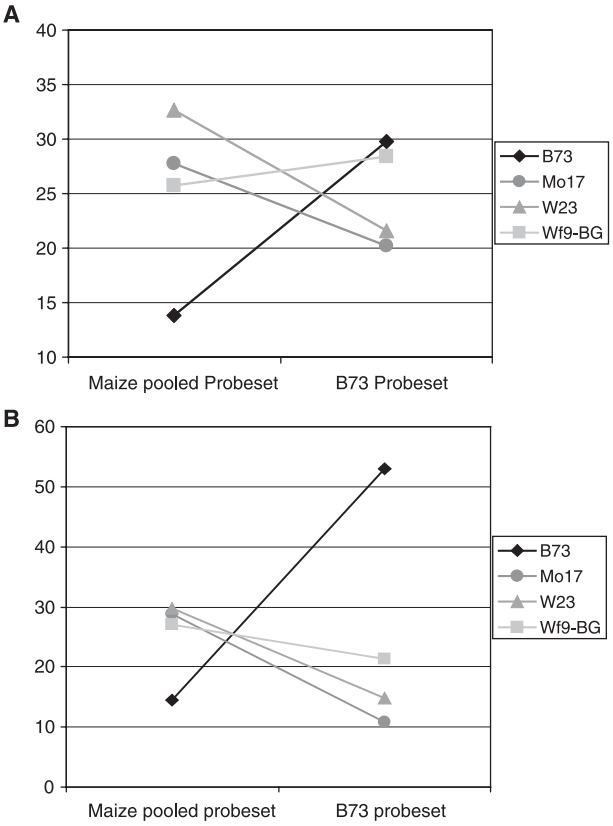

Figure 1 (A) Percentage of probesets in each line that display the highest signal estimate (relative to the other three lines, although not necessarily significantly higher). (B) Percentage of probesets that display a significantly higher signal estimate in one line relative to all the others. From probesets derived from the 'Maize pooled probeset' and 'B73 probeset', based on results from the Probe Logarithmic Intensity Error Estimation (PLIER) algorithms.

DNA sequence represented in the array are unique to certain lines, and more so when only sequences from a single line are used to design the probeset.

Probe and maize inbred line interactions

To identify sources of variation that could contribute to variation in probe intensities and probeset signal estimates, we applied ANOVA to the raw data measured in each probe of the array. Biological effects (i.e. differences in gene expression between maize inbred lines) were partitioned from technical sources of variation, including individual probe effect and interactions between probe and maize inbred lines (Figure 2). The former produces unreliable probeset signal estimates that may be associated with systematic errors (e.g. poor probe design, sequence errors in the reference genome or sequence polymorphism between the reference genome and samples), and the latter can cause interactions between individual probes and genotypes. Using ANOVA, we identified significant probe effects for all genes, verifying that not all probes 'behave' equally for any one given gene. Significant interactions between probes and maize inbred lines were detected for 36 196 probes (14.4% of the total) when comparing signal estimates detected in individual probes between maize lines (experiment-wise $\alpha = 0.05$ after Bonferroni correction). For 5118 probesets, at least one probe showed significant interaction with a maize inbred line. For half of these probesets, five or more probes out of the 30 were affected (Figure 3). The majority of these probes were designed on the basis of gene sequences from the reference genome B73 (3018 of 5118). The number of probes per probeset that interacted significantly with maize inbred lines was also substantially larger in the B73 probeset (median of six) than in the Maize pool probeset (median of three).

After the identification of significant probes by line interactions, we repeated the analysis excluding unreliable probes, with the expectation that the bias in gene expression estimates could be removed. Excluding these probes did not produce a completely unbiased distribution of the data, suggesting that polymorphisms remain undetected by our approach, as observed by the significant proportion of false negatives (one-third of probes with polymorphisms). Less stringent *P* value cut-offs, defined by 0.1%, 1% and 5% false discovery rates (Storey and Tibshirani, 2003), were also evaluated, but produced a substantial increase in the number of false positives. As a result, we could not identify an optimal significance threshold that concurrently avoided Type I and Type II errors. Correction of bias is not likely to be improved by the analysis of a larger number of probes per gene, as the proportion of probes affected by SNPs is likely to remain unchanged, as will the overall biasing effect.

Identification of polymorphic sites

Probe-by-line effects suggest that the DNA sequences of the maize inbred lines have properties that result in an interaction with specific probes in the array. A negative interaction may indicate polymorphisms unique to one or more maize inbred lines. To evaluate this hypothesis, we amplified genomic DNA by polymerase chain reaction (PCR), sequenced the products and contrasted the DNA sequence of a set of 11 genes (Table 1) from all four lines (B73, Mo17, W23 and Wf9-BG) relative to the probe sequence on the array. A total of 3852 base pairs was characterized from each line, covering 256 probes. Insertions, deletions and SNPs were identified in 88 loci (Table 1). In several cases, one or more polymorphisms were identified throughout the length of a single probe, and

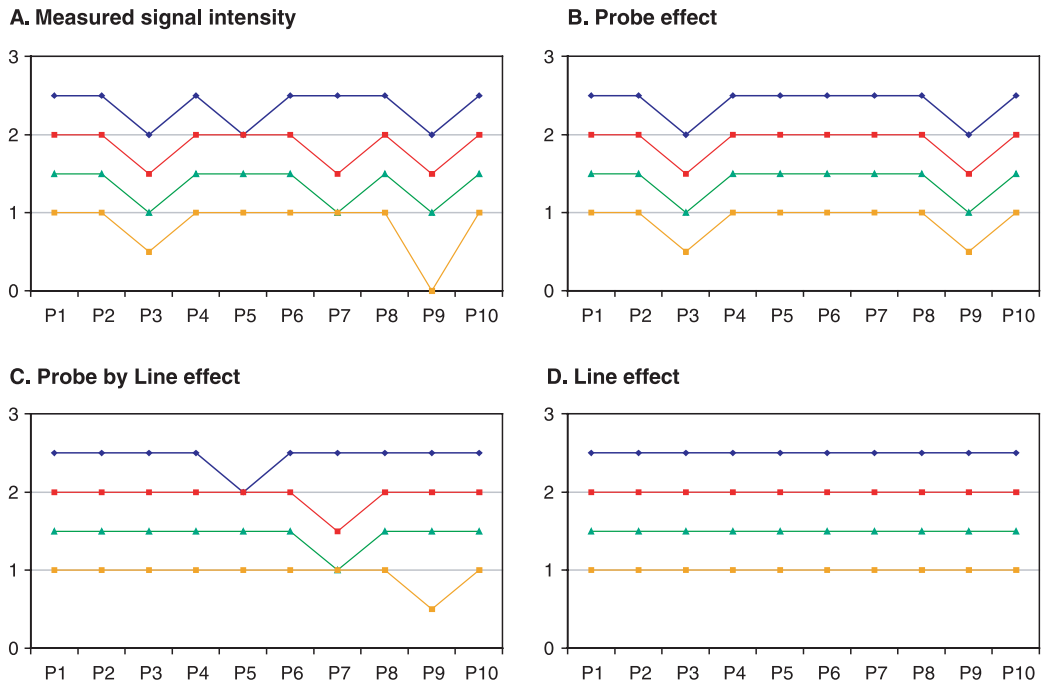

Figure 2 Partitioning the variance that contributes to signal intensity measurements. (A) Hypothetical signal intensity (*y*-axis) measured in a set of 10 probes (*x*-axis: P1–P10) representing one gene in four maize lines (blue, red, green and yellow lines). (B) Probe effects are detected for P3 and P9, where the signal intensity is consistently lower relative to the other probes in all maize lines. (C) Probe-by-line effects (i.e. blue, P5; red and green, P7; yellow, P9) indicate that specific probes interact negatively with specific lines. (D) After excluding probe and probe-by-line effects, the source of variation remaining is that associated with differences in transcript abundance between lines.

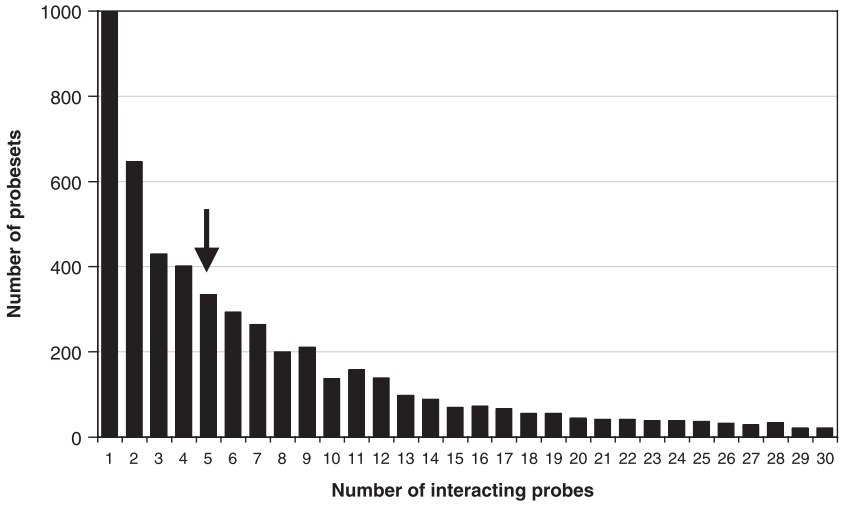

Figure 3 Probe-by-maize line interactions. Number of probesets (*y*-axis) in which one or more probes (*x*-axis) were detected as interacting significantly between maize lines. For one-half of the genes, five or more probes interact significantly with lines (median = 5, arrow).

some probes overlapped to such a degree that a single polymorphism was represented in up to three different probes (Figure 4, locus C). A total of 109 probes contained one or more polymorphisms in at least one maize inbred line.

With the location of DNA sequence polymorphisms known, we proceeded to evaluate whether DNA sequence variation between maize inbred lines translated into statistically significant differences between the probe-by-line effect estimated for the four lines. At a stringent Bonferroni cut-off value (experiment-wise P value < 0.05), significant differences were detected for 64 of the 109 probes (59%), indicating that more than one-third of the known polymorphisms remained undetected (false negative). A large proportion of polymorphisms located within the first five bases of the probe (positions 1–4 and 21–25) could not be detected (59%); in contrast, polymorphisms located within the five intermediate

Table 1 Properties of probesets characterized for genetic diversity and polymorphism detection

	Zm3.730	Zm3.4223	Zm3.1373	Zm3.5516	Zm3.5548	Zm3.7548	Zm3.7002	Zm3.5828	Zm3.15895	Zm3.16472	Zm3.5801	Total
Number of probes	23	24	20	23	20	26	24	27	27	21	21	256
Probe extension (bp)	412	388	288	331	264	400	373	437	384	291	284	3852
Number of polymorphisms	10	4	7	15	1	6	5	10	19	7	4	88
Probes including polymorphisms	10	7	10	14	1	7	7	13	20	13	7	109
Polymorphisms detected	5	2	8	6	1	2	7	13	7	8	5	64
False positives	0	0	0	0	4	0	0	0	0	0	0	4
False negatives	5	5	2	8	0	5	0	0	13	5	2	45

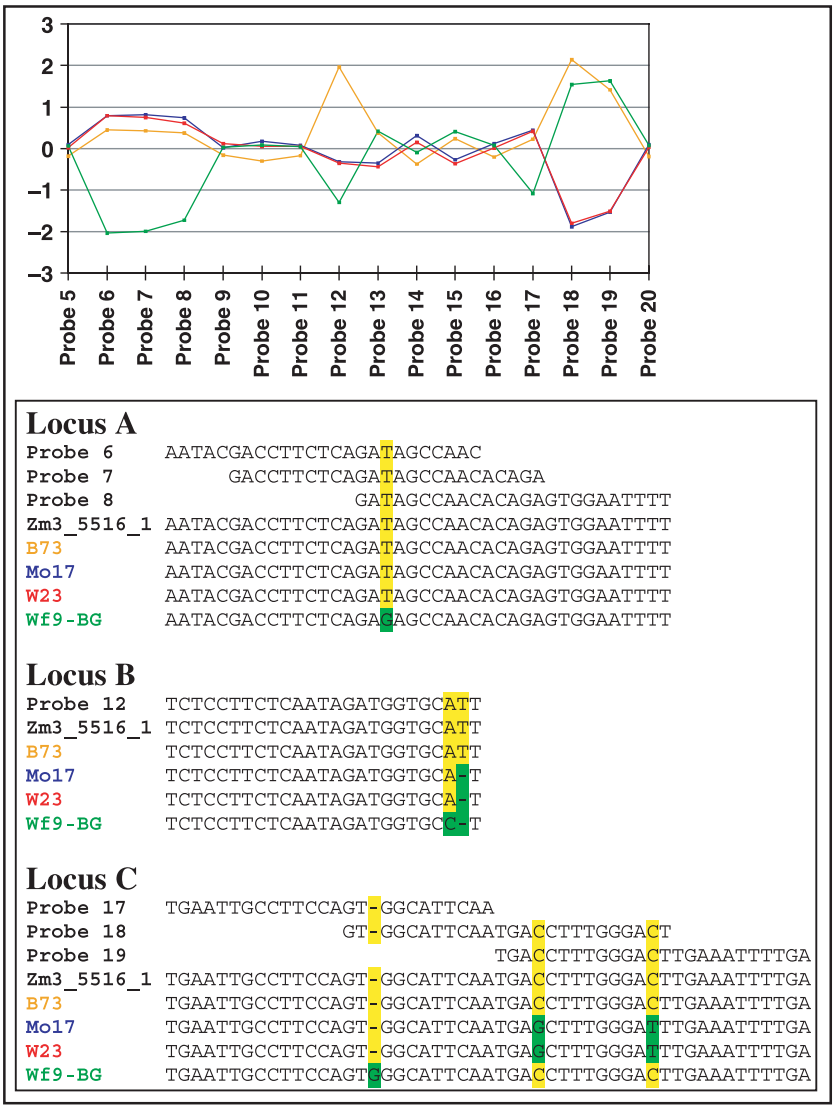

Figure 4 Contribution of polymorphisms to the signal detected in DNA array probes. Least-square means estimate of the probe-by-line effects (log2) in four maize inbred lines (yellow, B73; blue, Mo17; red, W23; green, Wf9-BG), relative to the probe mean across all lines (*y*-axis), for the gene Zm3.5516. The subset of probes 5–20 is represented (*x*-axis). Locus A: a T/G transversion on locus A affects probes 6, 7 and 8. Locus B: two polymorphisms in linkage equilibrium generate three haplotype classes that can be distinguished in probe 12. Locus C: probe 18 partially overlaps with probes 17 and 19, grouping the MIL into three distinct probe haplotype classes. For probe 17, an insertion (G) in Wf9-BG separates it from the other lines. Probe 19 is affected by two single nucleotide polymorphisms (SNPs) in linkage disequilibrium that separate the lines Mo17 and W23 from B73 and Wf9-BG. Probe 18 has a more complex pattern that results from the combination of the effect of SNPs and insertion that were identified in the probes 17 and 19, creating three haplotype classes.

positions (position 11–15) remained undetected in only 31% of the probes, as observed previously (Ronald *et al*., 2005; Rostoks *et al*., 2005). The position effect is evident in overlapping probes that contain a polymorphism in common, but the shared polymorphism is located in different regions of each probe (data not shown). For some genes, existing polymorphisms could not be detected because of high standard error associated with the line-by-probe effect estimates (e.g. Zm3.7548). The inability to detect significant differences in probe-by-line effect between maize inbred lines may be largely attributable to low heritability of gene expression for specific genes, thus adding variation to the detected probe intensities and probeset signal estimates and diminishing the power to separate the sources of variance. Technical issues such as cross-hybridization may also contribute to this variation, as poor probe design can result in multigenic contributions to probe intensity variation, thereby generating a complex signal estimate phenotype where genotype, probe and their interacting effects cannot be deconvoluted. This problem could potentially be alleviated during probe design when the current genome sequencing project for the B73 inbred line is completed. A reference genome will provide a substrate to compare possible cross-hybridization *in silico*.

A confounding issue in maize is that lines differ in the copy number of some genes. Such duplicated (or higher copy number genes) may be closely or distantly related. Variable copy number of genes with unknown levels of polymorphism could result in comparisons between non-homologous gene copies and, for some lines, in comparisons between the transcripts of one locus and nearly identical transcripts from two loci in another line. Some undetected polymorphisms located within the internal part of the probe could not be detected; we suggest that this may occur because repeats (homopolymers) are present, or there is the potential for secondary structure formation by probe and target (Karaman *et al*., 2005). For instance, probe 8 in Zm3.15895 and probes 13 and 17 in Zm3.5516 (Table 1) contain polymorphisms that were not detected, but, in each case, there is a strong potential for secondary structure formation (Gibbs free energy value $\Delta G < -3.0$ kcal/mmol).

In comparing sequences between maize inbred lines with the reference on the array, we detected several probes with two or more polymorphisms in linkage equilibrium, therefore generating three distinct 'probe haplotype' classes (e.g. loci B and C, Figure 4). Here, distinguishing the presence or absence of a mutation is not uniquely a test of the presence or absence of a signal relative to a control that represents the reference genome, as multiple levels may be present representing the combinatorial effect of several polymorphisms. This latter phenomenon results in haplotype grouping of genotypes that may be distinguishable. Probe haplotype classes could be defined statistically in some, but not all, cases.

Although false negatives were detected in roughly one-third of the probes in which polymorphisms were known to exist, only four false positives (i.e. significant differences between lines for a specific probe that could not be justified by the presence of a polymorphism) were detected in the set of 11 genes that were sequenced from all four maize inbred lines. A previous report (Rostoks *et al*., 2005) has suggested that 40% of discovered polymorphisms may be false positives. The lower proportion of false negatives discovered in our limited sample of genes (11), compared with those evaluated by Rostoks *et al*. (2005), is likely to reflect the more conservative threshold for assigning significance (Bonferroni vs. false discovery rate) and the properties of the experiment and the array (such as 30 vs. 16 probes) used in our study. It is also apparent that, depending on the overall level of polymorphisms in a gene, there will be a variation in the probability of detecting false positives or false negatives (Table 1). For an experiment-wise α of 5% (Bonferroni corrected), we identified 36 196 probes for 5118 probesets that showed significant differences between two or more maize lines. As a result, these probes may probably serve as individual molecular markers that could be used for genotyping. Indeed, the true number of useful 'scorable probes' for genotyping is perhaps much larger, as the correction for the multiplicity of tests is overly conservative because of the correlated nature of the data (several probes comprise a common polymorphism).

Discussion

In all microarray gene expression experiments, control of the sources of systematic variation is important. Bias from genetic diversity is just one of many systematic sources (e.g. method of sampling, biochemical methods to create labelled targets, array processing, analytical treatments of the data) that require consideration and careful inspection with regard to the severity of effects. Although a probeset signal estimate can display statistically significant systematic variation arising from polymorphism effects, it is appropriate in any experimental analysis to judge whether the systematic effect is significant relative to the biologically induced changes in transcript level being measured. As we have shown, the percentage of probeset signals that vary because of polymorphism can be large, as expected from the magnitude of signal estimate changes resulting from genetic variation. A comparison of the average signal detected in samples from the maize line B73 in the Maize pool probeset showed that

the signal was 9% lower than overall (across all lines, after chip normalization), and was only 0.5% (Wf9-BG) to 4% (Mo17 and W23) higher in the other lines. Conversely, in the B73 probeset, the average signal detected in the B73 lines was 10% higher than in the other maize lines, which displayed practically the exact opposite trend from that observed in the Maize pool probeset. In a preliminary experiment with maize landraces from Mexico and South America, an even more drastic impact of polymorphism was observed, in that fewer PM probes hybridized successfully (data not shown). These lines are uncharacterized at the DNA sequence level, but are predicted to be significantly more distant from B73 than the three inbred lines used in the primary study.

Multispecies sequence divergence has been shown to have a large effect on the hybridization signal in microarray experiments (Hsieh *et al*., 2003; Close *et al*., 2004; Gilad *et al*., 2005; Cho *et al*., 2006). In this study, we have demonstrated that this confounding effect can also be significant within an individual species. Our results are in sharp contrast with a study carried out in *Arabidopsis* (Kliebenstein *et al*., 2006), in which the effect of single-feature polymorphisms on gene expression estimates was considered to be negligible. Substantial differences in the levels of polymorphism in *Arabidopsis* and maize may explain the different conclusions. Polymorphisms are detected in *Arabidopsis* exons ($\theta_{exon} \approx 0.0025$; Nordborg *et al*., 2005) at a frequency that is approximately 1/3 of that observed in maize ($\theta_{coding} \approx 0.0072$; Tenaillon *et al*., 2001). At these mutation rates, two randomly selected coding sequences vary, on average, once every 400 bases in *Arabidopsis* and once every 139 bases in maize. Therefore, for a random sample of four *Arabidopsis* accessions and four maize inbred lines, a 25-mer probe is expected to contain a polymorphic site with a probability of 11% and 28%, respectively (see 'Experimental procedures' for calculation details). Even considering only polymorphisms within the 'internal' 15 bases, whose hybridization kinetics are known to be more highly affected by sequence mismatches (Ronald *et al*., 2005; Rostoks *et al*., 2005), the number of probes affected would be 18% in four random maize haplotypes, but only 7% in *Arabidopsis*. Therefore, it is not surprising that such a strong effect of polymorphisms is observed in our study relative to that in the analysis by Kliebenstein *et al*. (2006). In addition, the maize estimates presented here do not consider indels which, in our small sample of maize genes, represented 40% of all polymorphisms (data not shown). In summary, polymorphisms would be expected to significantly affect hybridization performance, on average, in approximately five to eight probes (18%–28% for a probeset of 30 probes) in our study with four maize lines, or more if indels were considered. Therefore, our experimental observation that, on average, approximately five probes show significant probe-by-line interaction (Figure 3) is likely to be an underestimate.

Different approaches may be applied to avoid the effect of polymorphisms on gene expression estimates. DNA sequence diversity could be taken into consideration in the array probe design if sequence information from a sufficiently large number of genetically diverse genotypes was available. Typically, genomic or coding (EST) DNA sequence information is available for one or few genotypes in any given species, but novel, very-high-throughput DNA sequencing methods may rapidly change this scenario. Extensive SNP discovery projects are also being carried out in some of the most genetically diverse plant species, including maize and pine, and the sequence diversity information can be used to support the design and application of species-wide reliable probes. Sequencing the most common haplotypes of genetically diverse species, such as maize, may lead to better array designs, avoiding polymorphic regions but still maintaining the capability to discriminate between gene family members.

More simplistic approaches of correcting for the effect of polymorphisms on gene expression estimates are to exclude probes with a significant probe-by-line interaction from the analysis, or to account for probe-by-line effects as a covariate in a mixed-model ANOVA. These methods did not produce a completely unbiased distribution of our data, suggesting that neither is sufficient for correcting for polymorphisms. An alternative is the application of highly stringent thresholds to define differential regulation between maize lines. This would probably minimize the biasing effect of polymorphisms, as the effect of SNPs and indels is expected to be less important than that of transcript abundance over all probes. A limitation of this approach is that genes with high genetic diversity – for which most probes comprise polymorphisms – may still be incorrectly defined as differentially expressed (false positives), and the higher stringency will exclude genes that may be biologically relevant from further analysis (false negatives). It is important to note that experiments involving single maize inbred lines under different treatments may circumvent the bias, as individual genes will be equally affected across treatments.

Biases in gene expression analysis of species with high genetic diversity may also be relevant when long oligomer (> 50-mer) and cDNA-based microarrays are used. Gene expression estimates, even when compared between species with a high level of similarity, such as primates, and using cDNA-based microarrays, have been shown to be highly biased towards falsely identifying higher expression in the reference sequence species (Gilad *et al*., 2005). Although

studies involving multiple long oligonucleotide probes per gene have not been carried out to our knowledge, the frequency of nucleotide polymorphisms detected in some highly diverse species signifies that they may frequently contain three or more SNPs and indels per probe, which may cause bias issues.

Although the probe-by-line effect represents a confounding source of variation for the identification of differentially expressed genes, it may also be applicable to SNP detection and genotyping. Genotyping by means of hybridizing labelled genomic DNA to DNA arrays has been well documented in *Arabidopsis* (Borevitz *et al*., 2003), yeast (Winzeler *et al*., 1998) and bacteria. Array genotyping of species with larger genomes, such as maize (2500 Mbp), barley (5000 Mbp) and gymnosperms (> 20 000 Mbp), however, may prove more challenging, because the single-copy, transcribed portion of the genome comprises a proportionally smaller fraction of the total genomic DNA as the *C* value increases. Methods of genomic DNA fractionation (e.g. methyl filtration) can be used to reduce this complexity, but their efficacy has not yet been reported for genotyping purposes. In this context, the use of cRNA hybridization may provide a better alternative for genotyping when using short oligonucleotide expression arrays (Cui *et al*., 2005; Rostoks *et al*., 2005).

Experimental procedures

Affymetrix Maize CornChip0

The Maize CornChip0 comprised 8403 probesets, each represented by a set of 30 PM and 30 MM probes. Probesets were designed from B73 (4050) or from a contig consensus of mixed maize line DNA sequences (4353), and were based on EST singletons (2811), EST clusters (4295) and other non-EST-derived cDNA sequences (1297).

Maize lines, experimental design and transcript profiling

Total RNA was extracted from the third, fourth and fifth leaves collected from 6–10 glasshouse-grown plants from the lines B73, Mo17, Wf9-BG (T-cytoplasm) and W23. RNA was isolated and purified using a hot (60 C) phenol/guanidine thiocyanate method (Caldo *et al*., 2004), and further purified using an RNeasy Midi Kit (Qiagen, Valencia, CA, USA). Plant growth and tissue collection were conducted in three independent biological replications in a randomized design. Total RNA was labelled according to the Affymetrix Expression Analysis Manual, and hybridized to two CornChips (technical replicates). Six replicates (three biological, two technical) of each line were hybridized to the CornChips.

Target synthesis and array hybridization

Probe synthesis, labelling and hybridization protocols were followed as described in the Affymetrix Manual (Affymetrix) and performed at the Iowa State University Microarray Core facility (http://www.biotech.iastate.edu/facilities/genechip/Genechip.htm). Ten micrograms of purified RNA with a 260/280 ratio of 2.0 were employed for cDNA synthesis using the One Cycle Kit (Affymetrix). Double-stranded cDNA was purified using the Gene Sample Cleanup Module, and 6 L of purified cDNA was employed to generate a biotinylated cRNA target using the GeneChip® IVT Labeling Kit. Labelled cRNA was purified using the Affymetrix Sample Cleanup Module, and the concentration of cRNA was adjusted on the basis of the total RNA used as starting material. Twenty micrograms of cRNA at a final concentration of 0.5 g/ L were fragmented in 5 × Fragmentation Buffer at 94 C for 35 min. The quality of cDNA, cRNA and fragmented cRNA was verified at each step on an Agilent 2100 Bioanalyser equipped with an RNA Nano LabArray (Agilent Technologies, Palo Alto, CA, USA). Fifteen micrograms of fragmented cRNA were used to make each hybridization cocktail and an equivalent of 10 g was hybridized to an array. Hybridization was performed at 60 C for 16 h in an Affymetrix Hybridization Oven model 640, arrays were washed and stained with streptavidin–phycoerythrin using the fluidics protocol EukGE-WS2v5 in the Affymetrix GeneChip® Fluidics Station 450, and stained arrays were immediately scanned with a GeneChip® Scanner 3000. The raw data are deposited in PLEXdb (*Pl*ant *Ex*pression *D*ata*b*ase; http://plexdb.org/) (Accession ZM1: CornChip0-Maize Pilot GeneChip Array) and the GEO database (Platform: GPL3618, Samples: Series: GSE4663, GSM105208–GSM105231, including supplementary CEL files).

Probe level analysis

Maize CornChips were scanned according to the manufacturer's instructions. The GeneChip® Operating Software GCOS (Affymetrix) and the Expression Intensity Explorer (Affymetrix Developer Network-Affymetrix Tools) were used to export PM and MM intensities into a text file.

These data were then analysed using a two-step strategy outlined previously (Chu *et al*., 2002; Hsieh *et al*., 2003), employing exclusively the PM probe data. Inclusion of the MM probe information has typically been found to reduce signal estimate bias at the expense of adding noise to the

data, decreasing the statistical power in detecting differences between variable levels (Chu *et al*., 2002; Hsieh *et al*., 2003). Use of the MM probe to remove probe-by-genotype effects was employed only when the PLIER and MAS5 approaches were applied. Individual probe measurements were first centred relative to the array mean by subtracting the log2-transformed mean value of the probes on the array. For the analysis of the commercial inbred lines, we proceeded by analysing each probeset individually in a mixed ANOVA in which line effects [B73, Mo17, Wf9-BG and W23; three degrees of freedom (d.f.)], probe effects (30 PM probes per gene; 29 d.f.) and line-by-probe effects (87 d.f.) were evaluated and least-square means were estimated. For all factors in the model (maize line, probe and probe-by-line), we evaluated the presence of significant variation using a test 3 for fixed effects. The model also included a technical replication as a random effect. All pair-wise comparisons (*t*-test) were carried out to contrast probe-by-line effects between maize inbred lines and to define potential polymorphisms between lines.

The maize inbred line effects were estimated from the output of Affymetrix PLIER and MAS algorithms, which generate signal values for each gene and DNA array by combining the information from PM and MM probes. For MAS signal estimates, a single point linear normalization was performed, whereas, for PLIER, quantile normalization was applied to the data set. The model was similar to that applied previously, except that probe effects and their interactions were not included, as PLIER and MAS combine their information in one value. All the analyses were carried out in SAS (SAS Institute, Cary, NC, USA).

Calculation for the probability of polymorphism presence

The number of segregating sites for four maize lines was estimated as follows: (i) by defining the probability of a nucleotide being polymorphic by multiplying theta (θ) by the corrected sample size (1/1 + 1/2 + 1/3); and (ii) by estimating the probability of non-occurrence of a polymorphism in a 25-mer probe [(1 – probability of a nucleotide being polymorphic)^{25}]. The probability that a probe contains a polymorphism can then be estimated by 1 – probability of non-occurrence of a polymorphism in a 25-mer probe.

DNA sequencing, alignment and polymorphism detection

Primer pairs were defined based on the original sequences used for probe design in the Affymetrix Maize CornChip0 (Supplementary Appendix S1). After PCR amplification in B73, Mo17, Wf9-BG and W23, the PCR products were purified by ethanol precipitation, resuspended in formamide, denatured and sequenced in an Applied Biosystems 3730xl DNA Analyser at the Cornell University Sequencing Facility. Trace files were visually inspected, aligned and polymorphisms were identified in BioEdit/BioAlign (www.mbio.ncsu.edu/BioEdit). The set of 11 genes reported in this study was selected on the basis of the following: (i) availability of a high-quality sequence for primer design; (ii) ability to design primers to PCR amplify regions that comprise at least 20 probes per gene; (iii) ability to produce specific, high-quality PCR amplifications in B73, Mo17, Wf9-BG and W23; (iv) ability to display significant probe-by-line interaction effects (test 3 for fixed effects); and (v) ability to produce high-quality sequences in the four maize lines after PCR amplification and sequencing.

References

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E. and Chory, J. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res*. **13**, 513–523.

Caldo, R.A., Nettleton, D. and Wise, R.P. (2004) Interaction-dependent gene expression in Mla-specified response to barley powdery mildew. *Plant Cell*, **16**, 2514–2528.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G.Q. and Lander, E.S. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet*. **22**, 231–238.

Cho, S.H., Garvin, D.F. and Muehlbauer, G.J. (2006) Transcriptome analysis and physical mapping of barley genes in wheat–barley chromosome addition lines. *Genetics*, **172**, 1277–1285.

Chu, T.M., Weir, B. and Wolfinger, R. (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math. Biosci*. **176**, 35–51.

Close, T.J., Wanamaker, S.I., Caldo, R.A., Turner, S.M., Ashlock, D.A., Dickerson, J.A., Wing, R.A., Muehlbauer, G.J., Kleinhofs, A. and Wise, R.P. (2004) A new resource for cereal genomics: 22K barley GeneChip comes of age. *Plant Physiol*. **134**, 960–968.

Cui, X.P., Xu, J., Asghar, R., Condamine, P., Svensson, J.T., Wanamaker, S., Stein, N., Roose, M. and Close, T.J. (2005) Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit. *Bioinformatics*, **21**, 3852–3858.

Gilad, Y., Rifkin, S.A., Bertone, P., Gerstein, M. and White, K.P. (2005) Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res*. **15**, 674–680.

Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N.P., Weder, A., Cooper, R., Lipshutz, R. and Chakravarti, A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet*. **22**, 239–247.

Hsieh, W.P., Chu, T.M., Wolfinger, R.D. and Gibson, G. (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics*, **165**, 747–757.

Karaman, M.W., Groshen, S., Lee, C.C., Pike, B.L. and Hacia, J.G. (2005) Comparisons of substitution, insertion and deletion probes for resequencing and mutational analysis using oligonucleotide microarrays. *Nucleic Acids Res.* **33**, e33.

Kliebenstein, D.J., West, M.A.L., van Leeuwen, H., Kim, K., Doerge, R.W., Michelmore, R.W. and St. Clair, D.A. (2006) Genomic survey of gene expression diversity in *Arabidopsis thaliana*. *Genetics*, **172**, 1179–1189.

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

Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H.G., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N.A., Shah, C., Wall, J.D., Wang, J., Zhao, K.Y., Kalbfleisch, T., Schulz, V., Kreitman, M. and Bergelson, J. (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *Plos Biol*. **3**, 1289–1299.

Ronald, J., Akey, J.M., Whittle, J., Smith, E.N., Yvert, G. and 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, J.O., Hedley, P.E., Russell, J., Mudie, S., Morris, J., Cardle, L., Marshall, D.F. and Waugh, R. (2005) Single-feature polymorphism discovery in the barley transcriptome. *Genome Biol*. **6**, R54.1–10.

Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genome wide studies. *Proc. Natl. Acad. Sci. USA*, **100**, 9440–9445.

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F. and Gaut, B.S. (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp *mays* L.). *Proc. Natl. Acad. Sci. USA*, **98**, 9161–9166.

Winzeler, E.A., Richards, D.R., Conway, A.R., Goldstein, A.L., Kalman, S., McCullough, M.J., McCusker, J.H., Stevens, D.A., Wodicka, L., Lockhart, D.J. and Davis, R.W. (1998) Direct allelic variation scanning of the yeast genome. *Science*, **281**, 1194–1197.

Supplementary material

The authors have provided the following supplementary material, which can be accessed online alongside the article at http://www.blackwell-synergy.com.

Appendix S1 EST and primer pair sequences used for amplification and sequencing for SNP analysis in the four maize inbred lines.