The Plant Genome
Original Research

Evaluation of Target Preparation Methods for Single-Feature Polymorphism Detection in Large Complex Plant Genomes

Michael Gore,* Peter Bradbury, René Hogers, Matias Kirst, Esther Verstege, Jan van Oeveren, Johan Peleman, Edward Buckler, and Michiel van Eijk

M. Gore, Dep. of Plant Breeding and Genetics, Institute for Genomic Diversity, Cornell Univ., 175 Biotechnology Building, Ithaca, NY 14853; P. Bradbury, USDA-ARS, Cornell Univ., 741 Rhodes Hall, Ithaca, NY 14853; R. Hogers, E. Verstege, J. van Oeveren, J. Peleman, and M. van Eijk, Keygene N.V., Agro Business Park 90, P.O. Box 216, 6700 AE Wageningen, The Netherlands; M. Kirst, School of Forest Resources and Conservation, Univ. of Florida, Gainesville, FL 32610; E. Buckler, USDA-ARS and Dep. of Plant Breeding and Genetics, Institute for Genomic Diversity, Cornell Univ., 159 Biotechnology Building, Ithaca, NY 14853. M. Gore and P. Bradbury contributed equally to this work. Received 14 Feb. 2007. *Corresponding author (mag87@cornell.edu).

Abstract

For those genomes low in repetitive DNA, hybridizing total genomic DNA to high-density expression arrays offers an effective strategy for scoring single-feature polymorphisms (SFPs). Of the ~2.5 gigabases that constitute the maize (*Zea mays* L.) genome, only 10 to 20% are genic sequences, with large amounts of repetitive DNA intermixed throughout. Therefore, a target preparation method engineered to generate a high genic–to–repetitive DNA ratio is essential for SFP detection in maize. To that end, we tested four gene enrichment and complexity reduction target preparation methods for scoring SFPs on the Affymetrix GeneChip Maize Genome Array ("Maize GeneChip"). Methylation filtration (MF), C_0t filtration (CF), mRNA-derived cRNA, and amplified fragment length polymorphism (AFLP) methods were applied to three diverse maize inbred lines (B73, Mo17, and CML69) with three replications per line (36 Maize GeneChips). Our results indicate that these particular target preparation methods offer only modest power to detect SFPs with the Maize GeneChip. Most notably, CF and MF are comparable in power, detecting more than 10 000 SFPs at a 20% false discovery rate. Although reducing sample complexity to ~125 megabase by AFLP improves SFP scoring accuracy over other methods, only a minimal number of SFPs are still detected. Our findings of residual repetitive DNA in labeled targets and other experimental errors call for improved gene-enrichment methods and custom array designs to more accurately array genotype large, complex crop genomes.

Published in Crop Sci. 47(S2) S135–S148. Published 14 July 2007.
doi:10.2135/cropsci2007.02.0085tpg

677 S. Segoe Rd., Madison, WI 53711 USA

Modern cultivated maize (*Zea mays* L.) boasts more genetic diversity than any other domesticated grass, retaining on average more than two-thirds of the nucleotide diversity of its wild relatives (Gaut et al., 2000; Tenaillon et al., 2001; White and Doebley, 1999). Indeed, DNA sequences of any two maize inbred lines differ from one another at an estimated frequency of a single nucleotide polymorphism (SNP) per 70 bases (silent sites) (Tenaillon et al., 2001). Considering such high levels of nucleotide diversity and a genome roughly equivalent in magnitude to the human genome (Arumuganathan and Earle, 1991), this yields about 30 million segregating sites. Intragenic linkage disequilibrium (LD) rates decline to minimal levels within two kilobases (kb) for a genetically diverse sample of tropical and temperate maize inbred lines (Remington et al., 2001). Due to this rapid breakdown of LD in a highly variable genome, an estimated one

Abbreviations: AFLP, amplified fragment length polymorphism; CF, C_0t filtration; FDR, false discovery rate; Gb, gigabase; HAP, hydroxyapatite; HC, High-C_0t; indel, insertion–deletion; kb, kilobase; LD, linkage disequilibrium; LTR, long terminal repeat; Mb, megabase; MF, methylation filtration; MM, mismatch; PM, perfect match; RMA, robust multichip average; SFP, single-feature polymorphism; SNP, single nucleotide polymorphism; SPB, sodium phosphate buffer; ss, single-stranded.

million SNP markers are required for genomewide association studies.

Although the maize genome is a sizable ~2.5 gigabases (Gb), the vast majority consists of several classes of retroelements known as long-terminal repeat (LTR) retrotransposons (SanMiguel et al., 1996). Long terminal repeat retrotransposons are generally recombinationally inert, thereby confining most meiotic recombination to the gene-rich or low-copy-number regions of the maize genome (Fu et al., 2002, 2001; Yao et al., 2002). Association mapping approaches, which rely on historical recombination for resolving complex traits, require that these regions of active recombination be identified and tagged. Because gene expression microarrays consist of oligonucleotides (oligos) designed from the sequence of expressed genes, they offer one potentially powerful means of genotyping thousands of recombinationally active gene regions in parallel. The genotyping of sequence polymorphisms with an expression array is based on the concept that a perfectly matched target binds to an oligo probe or feature with greater affinity than a mismatched target (Borevitz et al., 2003; Singer et al., 2006). If an individual oligo feature on an expression array shows a significant and reproducible difference in hybridization intensity between genotypes or strains, it can serve as a polymorphic marker or single-feature polymorphism (SFP). The goal of this study was to test the feasibility of expression arrays for use in SFP detection in maize.

The efficacy of Affymetrix (Santa Clara, CA) expression arrays for permitting highly accurate scoring of SFPs has already been demonstrated in relatively small genomes such as ~4-megabase (Mb) bacteria (*Mycobacterium tuberculosis*) (Tsolaki et al., 2004), ~12-Mb yeast (*Saccharomyces cerevisiae*) (Winzeler et al., 1998), and ~135-Mb *Arabidopsis thaliana* (hereafter *Arabidopsis*) (Borevitz et al., 2003). Expression arrays hybridized with DNA have also been used to map genetic loci and dissect traits (Singer et al., 2006; Steinmetz et al., 2002; Werner et al., 2005; Wolyn et al., 2004). Such whole-genome hybridization, however, has had limited success for detection of SFPs in crop plants with larger, more complex genomes, such as ~5.2-Gb barley (*Hordeum vulgare* L.) (Rostoks et al., 2005) and ~2.5-Gb maize (Kirst and Buckler, unpublished data, 2004). Thus, a target preparation method based on gene enrichment or complexity reduction is needed to exploit this potentially powerful technology.

One reasonably effective strategy is to score SFPs with cRNA derived from the less complex mRNA fraction of barley and maize (Cui et al., 2005; Kirst et al., 2006; Rostoks et al., 2005). Using cRNA as a surrogate for genomic DNA, however, has several notable limitations, including a requirement for extensive replication (e.g., 6X in Kirst et al., 2006) and a need to sample multiple tissues due to spatial and temporal expression of genes (e.g., 3X of six tissue types in Rostoks et al., 2005).

Methylation filtration (MF) with the bacterial *McrBC* restriction-modification system and $C_o t$ filtration (CF) are two gene-enrichment technologies that have enabled a significant proportion of the maize gene space to be sequenced (Palmer et al., 2003; Whitelaw et al., 2003; Yuan et al., 2003). They yielded a four- to sevenfold enrichment in maize gene sequences compared to control libraries (Rabinowicz et al., 1999; Yuan et al., 2003). Methylation filtration exploits the differential methylcytosine patterns between genes and retrotransposons in plants. Unlike mammalian retrotransposons, those in plants are more heavily methylated than the rest of the genome (Rabinowicz et al., 2003; Rabinowicz et al., 2005). When plant retrotransposon DNA containing methylcytosine on one or both strands is preceded by a purine (G/A) residue (Raleigh, 1992; Sutherland et al., 1992), it is cleaved by McrBC, a novel type I GTP-dependent restriction endonuclease. This results in gene rich regions being digested much less frequently than retrotransposon blocks—a characteristic that has been used to clone and sequence the unmethylated portion (gene space) of genomes from several plant genera (Bedell et al., 2005; Palmer et al., 2003; Rabinowicz et al., 1999, 2005).

The principle underlying CF is based on the renaturation kinetics of DNA (Britten and Kohne, 1968) and has been used to differentially fractionate plant genomes according to copy number and base composition (Geever et al., 1989; Hake and Walbot, 1980; Peterson et al., 2002a; Yuan et al., 2003). Mechanically sheared genomic DNA is denatured and reassociated to a calculated $C_o t$ value, a product of nucleotide concentration and reassociation time (Peterson et al., 2002a). The unrenaturated genome fraction enriched for low-copy number and genic sequences (High-$C_o t$) is then cloned and sequenced, while the renaturated moderately (Medium-$C_o t$) and highly repetitive (Low-$C_o t$) DNA fractions are excluded (Peterson et al., 2002a; Yuan et al., 2003).

A final technique, amplified fragment length polymorphism (AFLP), uses the random distribution of restriction endonuclease recognition sites across a genome to make amplification libraries (Vos et al., 1995). By carefully selecting enzyme motifs and varying the number of selective bases in the amplification primers, it is possible to modulate both the number of unique, amplified fragments as well as genome complexity. Although standard AFLP

procedures are not biased to gene regions, different random pools of DNA can be preferentially amplified and genotyped on expression arrays by changing enzymes. Amplified fragment length polymorphism offers the additional advantage of being reproducible and amenable to high throughput processing.

Due to large amounts of repetitive, mobile DNA, the maize genome requires a target preparation method that offers both a high level of gene enrichment and accurate scoring of SFPs. The objectives of this paper are (i) to determine which target preparation method (CF, MF, mRNA, or AFLP) optimally enriches for gene sequences complementary to probe sequences on the Affymetrix GeneChip Maize Genome Array and (ii) to estimate SFP detection power for each target method.

MATERIALS AND METHODS

Sample and Array Specifications

To evaluate the effectiveness of several target preparation methods for detecting SFPs in large, complex plant genomes, we conducted an experiment to score SFPs in three diverse maize inbred lines. Iowa Stiff Stalk Synthetic line, B73; non-stiff stalk line, Mo17; and tropical lowland CIMMYT (International Center for Maize and Wheat Improvement) line, CML69 represent the three major subpopulation groups of maize inbred lines (Liu et al., 2003; Remington et al., 2001). The Affymetrix Gene Chip Maize Genome Array ("Maize GeneChip") has 17 555 probesets with 263 026 probe pairs for expression profiling 14 850 maize genes (13 339 unique). Of the 17 555 probesets, 17 477 have 15 probe pairs, while the remaining 78 probesets have 14 or less probe pairs. Each probe pair consists of a perfect match (PM) probe and mismatch (MM) probe. The PM probe has a 25-bp sequence that is identical to a specific target gene transcript, whereas the MM probe differs from the PM probe by a single nucleotide substitution at the central base position. Array probes are designed from the sequence of expressed maize genes available in NCBI's GenBank (up to 29 September 2004) and *Zea mays* UniGene Build 42 (23 July 2004) databases (http://www.affymetrix.com).

Target Synthesis and Array Hybridization

Total genomic DNA was extracted from powdered lyophilized leaf tissue using cetyltrimethylammonium bromide (CTAB) extraction buffer according to the protocol described by Saghai-Maroof et al. (1984). DNA was extracted in triplicate from a single genotyped tissue source; thus, all DNAs isolated from the same inbred tissue source are technical replicates.

The maize genome was methylation filtered using McrBC as previously described by Zhou et al. (2002), with minor modifications. McrBC fragments were generated by incubating 60 µg genomic DNA with 600 U of McrBC (New England Biolabs, Ipswich, MA) at 37°C for 8 h, followed by heat inactivation of the enzyme at 65°C for 20 min. McrBC fragments ranging in size from ~12 kb to less than 100 bp (data not shown) were separated on a low-melting 0.8% SeaPlaque Agarose gel (Cambrex Bio Science Rockland, Inc., Rockland, ME). Most unwanted, restricted methylated DNA migrated to positions below the 1-kb marker. Fragments ≥1 kb were excised from the gel and purified using the QIAEX II Gel Extraction Kit (QIAGEN, Valencia, CA), according to the manufacturer's protocols.

$C_o t$ filtration involved selecting the High-$C_o t$ (HC) single-stranded (ss)DNA fraction as described by Peterson et al. (2002a). In brief, 50 µg of genomic DNA was sheared to an average fragment size of 450 bp using a Misonix Sonicator 3000 (Misonix, Inc., Farmingdale, NY) with full power settings, for 24 cycles of 30 s of sonication and 1 min of cooling. Cations were removed using a Chelex ion-exchange column, followed by concentration and resuspension of the DNA in 0.5 *M* sodium phosphate buffer (SPB). DNA was transferred to capillary tubes, denatured in boiling water for 10 min, and allowed to renature to a $C_o t$ value of 262 *M*·s. A $C_o t$ value is the product of the sample's nucleotide concentration (moles of nucleotides per liter), its reassociation time in seconds, and a buffer factor based on cation concentration (Peterson et al., 2002a). Renatured DNA was then transferred to a hydroxyapatite (HAP) column (Bernardi, 1971) equilibrated with 0.03 *M* SPB. Finally, HC ssDNA was eluted by loading the HAP column with 0.12 *M* SPB.

Amplification of AFLP fragments was performed according to the protocol described by Vos et al. (1995), using 200 ng genomic DNA as starting material. Sequences of the *Taq*I adaptor were 5′-CTCGTAGACTGCGTAC-3′ and 5′-CGGTAC-GCAGTCT-3′, and sequences of the *Mse*I adaptor were 5′-GACGATGAGTCCTGAG-3′ and 5′-TACT-CAGGACTCA-3′. Sequences of the TaqI+A, MseI+C and MseI+G primers were 5′-GTAGACTGCG-TACCGAA-3′, 5-GATGAGTCCTGAGTAAC-3′ and 5′-GATGAGTCCTGAGTAAG-3′, respectively. Amplified fragment length polymorphism products were purified by standard sodium acetate–ethanol precipitation and dissolved in $T_{10}E_{0.1}$.

A total of 300-ng purified HC ssDNA, MF DNA, or purified AFLP product were biotin-labeled in triplicate using the BioPrime DNA labeling system (Invitrogen, Carlsbad, CA), as described by Borevitz

et al. (2003). Specifically, 60 µL 2.5X random octamer primers and 300 ng DNA were denatured in a total volume of 132 µL at 95°C for 10 min and cooled on ice to allow annealing of random primers. Next, 15 µL 10X dNTP/biotin-14-dCTP and 3 µL Klenow fragments were added for primer extension and incubated overnight at 25°C. Labeled fragments were purified by standard sodium acetate/ethanol precipitation and dissolved in 30 µL $T_{10}E_{0.1}$. For the labeled AFLP samples, a total of 15 µg *Taq*I+1(A)/*Mse*I+1(C) and 15 µg *Taq*I+1(A)/*Mse*I+1(G) from each sample were pooled and enough $T_{10}E_{0.1}$ was added to bring the final volume to 30 µL. The combination of these two AFLP +1/+1 samples was intended to represent an approximately 125-Mb fraction of the maize genome, which is almost equal in size to the *Arabidopsis* genome. These primer–enzyme combinations, however, are not optimized to specifically target gene regions.

Total RNA from homogenized frozen 4-wk-old leaf tissue was isolated using TRIZOL reagent (Invitrogen, Carlsbad, CA) and Qiagen RNAeasy Columns (QIAGEN, Valencia, CA) according to the manufacturers' protocols. Total RNA was isolated from harvested leaves of individual plants; thus, all RNAs isolated from a specific inbred are biological replicates. A total of 7 µg of each RNA sample was used for double-stranded cDNA synthesis and biotin-labeling of antisense cRNA, as described in the manual accompanying GeneChip Expression 3′-Amplification Reagents One-Cycle cDNA Synthesis Kit and One-Cycle Target Labeling Assay (Affymetrix, Santa Clara, CA). Finally, 15 µg biotin-labeled cRNA per reaction was supplemented with $T_{10}E_{0.1}$ to achieve a final volume of 30 µL.

Hybridizations on GeneChip Maize Genome Arrays (Affymetrix, Santa Clara, CA) were performed by an Affymetrix service station (ServiceXS, Leiden, The Netherlands), according to Affymetrix protocols. In total, 36 GeneChips were used in this study. Three technical replicates of CF, MF, and AFLP for each line were hybridized to 27 GeneChips, and three biological replicates of mRNA for each line were hybridized to 9 GeneChips.

GeneChip Quality Control

The scanned image of each GeneChip was visually inspected for spatial artifacts using the method Image of the *affy* package (http://www.bioconductor.org) in the freely available statistical package R (http://www.r-project.org; Ihaka and Gentleman, 1996). Standard Affymetrix quality control parameters for assessing arrays were checked and determined to be reasonably concordant with the manufacturer's recommendations (Gene-Chip Expression Analysis Data Analysis Fundamentals; http://www.affymetrix.com).

Pearson's correlations of raw PM probe intensities between arrays of the same target preparation method ranged from 0.95 to 0.99 within line, while between lines correlations were in the range of 0.85 to 0.95. Notably, our analysis revealed that one of the Mo17 line-CF replicates had low correlations (0.5–0.6) to the other CF lines and replicates. Therefore, we excluded this outlier array from all further analyses. The inbred line assignment for each GeneChip was further verified by analyzing the average Euclidean distance between standardized $\log_2$ probe intensities of 289 probesets. All quality control statistical analyses were performed using SAS (SAS Institute, Cary, NC). The PROC CORR and PROC DISTANCE statements were used to calculate correlations and distances, respectively.

Maize Sequence Validation Dataset

Methodology

A dataset for validation of detected SFPs was created from sequence alignments that matched the sequence of probes on the Maize GeneChip (Maize_probe_tab.txt; http://www.affymetrix.com). Specifically, the 25-bp nucleotide sequence of each PM probe was compared to a 25-bp sliding window of nucleotide sequence along all B73, Mo17, and CML69 sequence alignments in the Panzea database (http://www.panzea.org) (Zhao et al., 2006). The reverse complement of each PM probe sequence was also used to search Panzea. If an exact match between an alignment and PM probe sequence was identified for at least one of the lines, a 25-bp string initiated from the probe start position within the alignment was extracted for all three lines. All three extracted 25-bp strings were then aligned to the initial queried PM probe sequence. This allowed for the number of exact match nucleotides to be counted and the position of any SNPs within the string to be recorded. Any extracted string containing a gap (insertion or deletion) or ambiguous nucleotide was discarded. The resulting sequence dataset contained all B73, Mo17, and CML69 sequences from Panzea that exactly matched Affymetrix PM probes for at least one of the inbred lines, along with any corresponding mismatch sequences from the remaining lines.

Additional criteria were used to help ensure the quality of sequences in the SFP validation dataset. For example, many of the alignments included two sequencings of B73 and Mo17 for quality control. If the two B73 strings or the two Mo17 strings were not identical for any 25-bp nucleotide sequence, the sequence at that position was not used. Also, on rare occasion (<0.5%) one of the lines was found to have more than four SNPs when compared to the probe

sequence. Sequence at that location was excluded from the dataset, as these SNPs may have been caused by an alignment error rather than actual sequence variation.

Primary SFP Validation Dataset

The primary SFP validation dataset was used to calculate SFP detection power for each target preparation method. This validation dataset contains 38 259 sequences of 25 bp (~1 Mb) from B73, Mo17, and CML69 for 14 651 PM probes, of which 1620 probes (11%) detect one to four SNPs in at least one of the three maize inbred lines. There are a total of 1998 segregating sites (*S*), which translates to a $\theta_{PMprobe}$ estimate of 0.0014. The number of SNPs detected by a PM probe in each inbred line is as follows: B73, 453; Mo17, 1070; and CML69, 802. Of the 14 651 PM probes with available sequence data for a maize inbred line, there are a maximum of 32 511 pairwise probe comparisons, and 2677 (8.2%) of these involve a PM probe that detects at least one SNP—potentially leading to the detection of 2677 SFPs. The calculated SFP rate in this dataset for each inbred pairwise probe comparison is as follows: B73-CML69, 7.9% (742/9386); B73-Mo17, 8.3% (1128/13 631); and CML69-Mo17, 8.5% (807/9494). Consequently, with this dataset, we can detect at most 2677 SFPs with each target preparation method if all 14 651 PM probes are members of probesets called Present (detected) by the Affymetrix Microarray Suite version 5 (MAS5) algorithm (Liu et al., 2002) on all CF, MF, mRNA, or AFLP arrays.

The observed SNP diversity ($\theta_{PMprobe} = 0.0014$) in the primary SFP validation dataset is about 19% of the SNP diversity ($\theta_{PMprobe} = 0.0075$) reported by Kirst et al. (2006) when PM probes were used to genotype a diverse set of maize inbred lines. In Kirst et al. (2006), cRNA was hybridized to an 8K Maize CornChip0, which contains probes that were designed from the sequence of a limited number of maize genotypes (e.g., ~50% B73 sequence). Unlike the Maize CornChip0, probes on the Maize GeneChip were designed to be robust for multiple maize genotypes by masking polymorphisms identified in the expressed sequences of over 100 maize lines (http://www.affymetrix.com; verified 12 June 2007; Stupar and Springer, 2006). Therefore, probes on the Maize GeneChip were systematically designed to hybridize regions of gene transcripts with lower than average levels of nucleotide diversity and, as such, resulted in low rates of SNP detection in this study.

Secondary SFP Validation Dataset

The secondary SFP validation dataset was used to calculate SFP detection power in an unbiased manner. This secondary dataset, a subset of the primary SFP validation dataset, was constructed with only PM probes from probesets that were called Present by MAS5 on all CF, MF, and mRNA arrays. Amplified fragment length polymorphism was not analyzed with the secondary SFP validation dataset due to the low number of shared probesets called Present by MAS5 on AFLP arrays. The secondary SFP validation dataset contains 23 873 sequences of 25 bp (~0.6 Mb) from B73, Mo17, and CML69 for 9039 PM probes, of which 835 PM probes (9.2%) detect one to four SNPs in at least one of the three maize inbred lines. With the 9039 PM probes, there are 20 666 pairwise probe comparisons, of which 1409 (6.8%) could potentially detect an SFP.

Polymorphic Probeset Validation Dataset

We also investigated whether probesets (probeset level analysis) containing one or more polymorphic probes (polymorphic probesets) are detected with greater accuracy than SFPs (probe level analysis). A dataset for validation of detected polymorphic probesets was constructed using probesets for which all probe sequences and SNPs were known. In the SFP validation dataset described above, very few probesets had all 15 probes match a sequence in the Panzea database. To construct a dataset of probesets with no missing sequence data, we first identified probesets that were called Present by the MAS5 algorithm on all CF, MF, and mRNA arrays. Second, probesets with eight or more probes matching an alignment sequence were identified. Third, probes within those probesets that had no matching Panzea sequence were removed from the dataset. The resulting probeset validation dataset contained 289 probesets, each consisting of between 8 and 15 probes. Of these 289 probesets, a total of 109 (38%) contained at least one mismatch probe due to a SNP in one of the three lines and as such were defined as polymorphic.

Hybridization Data Preprocessing and Normalization

Raw CEL files were background corrected (robust multichip average [RMA]; Irizarry et al., 2003) and then normalized (quantiles; Bolstad et al., 2003). We found that processing the hybridization data with RMA and Quantiles resulted in equivalent or higher SFP detection power as that obtained with the spatial correction method described in Borevitz et al. (2003). MAS5 was used to remove probesets called Absent or Marginal (unreliably detected) before probe level analysis. Probesets were retained for further analysis if called Present (detected) for a method specific set of nine GeneChips (MF, mRNA, and AFLP) or eight GeneChips (CF). Robust multi-array average, quantiles,

and MAS5 methods of the *affy* package were performed in R.

Detecting SFPs in Hybridization Data

Single-feature polymorphisms were identified in preprocessed hybridization data using the two-step strategy mixed model as described in detail by Kirst et al. (2006). Analyzed datasets of background, normalized probe intensities were derived from probesets called Present that included at least one probe sequence in common with the SFP validation dataset. Each probeset was analyzed separately. The overall array mean for each array was subtracted from the $\log_2$ of the probe intensity. The following mixed model was fit to the resulting values in SAS:

$$I_{ijk} = L_i + P_j + a_{ik}(L_i) + \mathrm{L_i}^{*}P_j + e_{ijk}$$

where I_{ijk} = log2(probe intensity) for the ith maize inbred line for the jth probe on the kth array of that line less the array mean for the probeset; L_i = the effect of the ith line; P_j = the effect of the jth probe; $a_{ik}(\mathrm{L}_i)$ = the effect of the kth array of the ith line nested within line; $L_i{}^{*}P_j$ = the interaction effect for the ith line and the jth probe—represents SFPs; e_{ijk} = random error; i = 1, 2, 3, the number of lines; j = 1, ... , 15, the number of probes in a probeset; and k = 1, 2, 3, the number of arrays per line.

The data were analyzed using SAS PROC MIXED, fitting line and probe as fixed effects and array as a random effect nested in line using the following model statements:

```
model intensity = line probe line*probe;
random array/subject = line;
lsmeans line*probe/diff
```

The LSMEANS statement in SAS was used to generate pairwise comparisons between inbred lines at each probe with a *t* test of the null hypothesis that the difference was zero. A statistically significant non-zero value indicated a potential SFP. All pairwise *t* test comparisons were performed in one of two ways: using the standard error from the probeset as indicated in the model above (probeset error term *t* test) or assuming a constant error term from the complete array (array error term *t* test).

The SFP validation sets were used to confirm whether detected SFPs were true or false positives, thereby allowing for the estimation of detection power at empirically calculated false discovery rates (FDRs). To do this, comparisons between lines at each probe were first sorted by p-value. For each *p* value, the FDR was calculated as the number of comparisons with an equal or lower *p* value that were false SFPs divided by the total number of comparisons with an equal or lower *p* value. The power was calculated as the number of true SFPs with an equal or lower *p* value divided by the total number of true SFPs in the dataset. Calculations were performed for both the probeset error term *t* test as well as the array error term *t* test.

All R and SAS scripts, raw GeneChip data, sequences for validation set probes, and lists of identified SFPs are available on request. Raw GeneChip data will also be deposited in PLEXdb (Plant Expression Database; http://plexdb.org; verified 12 June 2007).

RESULTS

Probe Performance

More than 14000 PM probe sequences on the Maize GeneChip are an exact match to at least one of the three maize inbred line sequences in the Panzea database. In these cases, we expect PM probe signal intensity to be greater than that of the MM probe. However, there are many instances where the MM probe has higher signal intensity than the PM probe (MM > PM) despite the fact that the PM probe is known to be a perfect match for the target. Figure 1 shows the distribution of PM probe signal intensities and indicates the portion for which the MM signal exceeds the PM signal. The PM signal on AFLP and mRNA GeneChips is strongly skewed toward the lower end of the $\log_2$PM intensity range, while CF and MF exhibit a more normal distribution. As the signal intensity of PM probes on mRNA GeneChips increases, the proportion of MM > PM probe pairs drastically diminishes. In comparison, the proportion of probe pairs on MF and CF GeneChips where the MM probe signal intensity is greater than the PM probe is more uniform across the $\log_2$PM signal intensity range.

Analysis of PM and MM probe signal intensity data from mRNA GeneChips confirms that when gene expression levels are high, signal from target sequence overwhelms the noise from nontarget sequences cross-hybridizing to probes. The main problem with mRNA samples is that they contain transcripts from genes with low levels of expression (i.e., signal near background levels) that in many cases makes SFP detection difficult. In contrast, CF and MF samples most likely contain low to intermediate levels of repetitive DNAs that are spuriously annealing to probes across all PM intensity levels. On the other hand, the most important factor affecting AFLP samples is that they are not well represented by GeneChip probes.

Array Coverage of Gene Enrichment Methods

Because of the significant number of identified MM > PM probe pairs, we used the Affymetrix

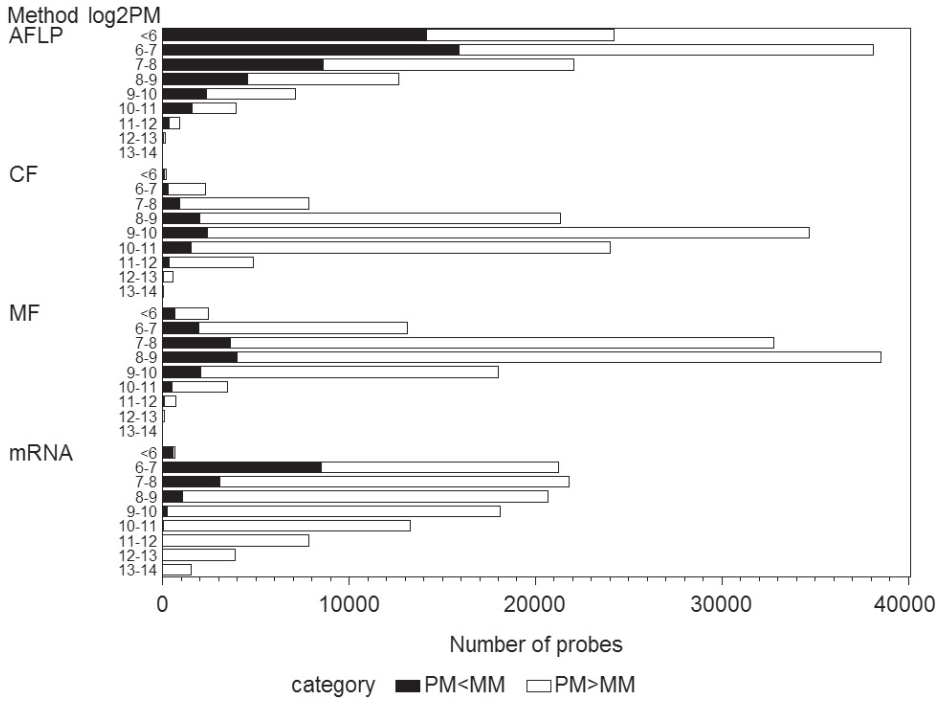

Figure 1. Frequency distribution of perfect match (PM) and mismatch (MM) probe pair signal intensity ratios. Probe pair signal intensity ratios are shown according to log2PM range for amplified fragment length polymorphism (AFLP), mRNA, $C_o t$ filtration (CF), and methylation filtration (MF).

Microarray Suite version 5 (MAS5) algorithm to filter hybridization data so that data for probesets unreliably detected could be eliminated. The MAS5 algorithm uses probe pair data in a Wilcoxon (1945) signed rank test to determine whether PM probes have a higher hybridization intensity signal than their analogous MM probes (Liu et al., 2002). Depending on the outcome of this test, one of three detection calls (Present, Absent, or Marginal) is assigned to each probeset. We performed a separate MAS5 analysis on each GeneChip. Hybridization data were maintained if probesets were called Present for each GeneChip in a target preparation method set, while data from probesets called Marginal or Absent were removed from further analyses.

Although the primary purpose for employing the MAS5 algorithm was to increase the ratio of true positive to false positive SFPs (i.e., decrease Type I error rate), this analysis also allowed us to calculate the total number of probesets called Present for GeneChips of each target preparation method. Because probes are designed from the sequence of expressed maize genes, the number of probesets called Present serves as a direct indicator of how well each method provides sequences complementary to probes on the Maize GeneChip. The number of probesets called Present by MAS5 differs substantially by target preparation procedure: AFLP, 646 (4%); mRNA, 9661 (55%); MF, 12 975 (74%); and CF, 14 895 (85%). $C_o t$ filtration and MF provide for a greater representation of complementary gene sequences than mRNA fractions isolated from a single tissue type (leaf) and specific developmental stage (V4-5). A larger portion of the maize gene space is sampled by CF and MF, while transcript presence and location are dependent on the temporal and spatial pattern of gene expression. Amplified fragment length polymorphism has more than 10-fold fewer Present calls, suggesting that the selected restriction enzymes (*Taq*I and *Mse*I) and amplification protocol substantially reduce maize genome complexity without highly enriching for gene fragments complementary to array probes.

Assessment of Power to Detect SFPs

To estimate SFP detection power afforded by CF, MF, mRNA, and AFLP, we first constructed a primary SFP validation dataset containing all B73, Mo17, and CML69 sequences from the Panzea database that matched to a PM probe sequence (see detailed description in "Materials and Methods" under "Maize Sequence Validation Dataset"). We determined that 1620 out of the 14 651 validation dataset probes should detect one to four SNPs (SNP probes) in at least one inbred line. The other 13 031 probes in the SFP validation dataset should not detect any SNPs when hybridized to target sequences from any of the three inbred lines (non-SNP probes).

Of the possible 32 511 pairwise probe comparisons between B73, Mo17, and CML69, there are 2677 comparisons that could potentially detect an SFP. The number of SNP and non-SNP validation dataset probes contained within probesets called Present by MAS5 was determined for each target preparation method (Table 1). The number of detected SNP and non-SNP probes shared with the primary SFP validation dataset is highest for CF and MF, which reflects their overall success in enriching for genes represented as probes on the array. Subsequently, we calculated the total number of potential SFPs that could be identified through pairwise probe comparisons of all three lines with MAS5 detected SNP and non-SNP probes (Table 1). $C_o t$ filtration and MF provide for a greater representation of probes on the GeneChip and in the SFP validation dataset and, as such, have the potential to provide more opportunities to detect SFPs.

We applied a mixed model to background, normalized probe intensity data from all probes of probesets called Present that share at least one probe sequence in common with the SFP validation dataset. The mixed model accounts for line, probe, and probe-by-line effects, which are sources of variation in probe intensities and probeset signal estimates (Kirst et al., 2006). A significant negative interaction between a probe and one or more inbred lines suggests that at least one DNA sequence polymorphism is reducing the signal intensity of the probe. Significant probe-by-line effects were detected using pairwise comparisons of individual probe intensity estimates between inbred lines, and SFP detection power was calculated at 5, 10, 20, 30, and 40% FDRs for each target preparation method (Table 2).

Power to detect SFPs was calculated as the proportion of detected true positive SFPs (sequence confirmed) to the total number of expected SFPs (Table 1) at an empirically determined FDR. Single-feature polymorphism detection power for mRNA was calculated using the probeset error term; for the other three methods, it was calculated using the array error term. Single-feature polymorphism detection power at 5% FDR is almost negligible (1%) for both MF and CF. Methylation filtration does detect 284 confirmed SFPs at 10% FDR, while the number of confirmed SFPs detected by CF at higher FDRs exceeds all other methods. Single-feature polymorphism detection power of mRNA and MF are almost equivalent at FDRs of 20% and higher, but more SFPs are detected using MF by virtue of its greater probe coverage. Amplified fragment length polymorphism scores SFPs with more accuracy than the other target preparation methods, but the numbers of detected SFPs are far lower due to AFLP's inferior SFP detection potential with this particular GeneChip design. This low potential directly results from the primary amplification of non-genic random sequences. Interestingly, no additional SFP detection power is gained until 60% FDR with AFLP, as power is static at 45% from 5 to 40% FDRs. A likely explanation is that AFLP accurately scores all the SFPs for the few genes that it can at 5% FDR with limited cross-hybridization from other amplified targets.

Table 1. Single-feature polymorphism (SFP) detection potential of target preparation methods.

Method†	Non-SNP probes‡		SNP probes§		SFPs¶	
	No. probes#	%††	No. probes	%	No. SFPs‡‡	%§§
CF	12 197	94	1430	88	2429	91
MF	10 851	83	1202	74	2009	75
mRNA	9285	71	1019	63	1707	64
AFLP	229	2	26	2	38	1
Total	13 031	100	1620	100	2677	100

†CF, $C_o t$ filtration; MF, methylation filtration; AFLP, amplified fragment length polymorphism.

‡Non-SNP probes, primary validation dataset probes that should not detect a single nucleotide polymorphism (SNP) in B73, Mo17, and CML69.

§SNP probes, primary validation dataset probes that should detect anywhere from 1 to 4 SNPs in B73, Mo17, and/or CML69 (not all three).

¶SFPs, individual probes that should detect at least 1 SNP in B73, Mo17, and/or CML69 (not all three) and therefore have potential for being detected as polymorphic markers or single-feature polymorphisms (SFPs) in pairwise probe comparisons.

#Number of validation non-SNP and SNP probes that are contained within a probeset called Present by MAS5.

††Percentage of all non-SNP or SNP probes in the primary validation dataset that are members of probesets called Present by MAS5 on arrays of each target preparation method.

‡‡Number of pairwise probe comparisons that could potentially detect an SFP. These calculated numbers are based on the specific MAS5 detected non-SNP and SNP probes for each target preparation method and are cumulative across the three inbred lines (B73 vs. Mo17; B73 vs. CML69; or Mo17 vs. CML69). Not all inbred line pairwise comparisons were possible for some probes because sequence information was missing for one of the lines.

§§Percentage of all SFPs represented in the primary validation dataset that are detectable on arrays of each target preparation method.

The mixed model was also applied to a subset of the probe intensity data that consists of 1440 probesets called Present on all CF, MF, and mRNA GeneChips. All of the parsed probesets have one or more probe sequences in common with the secondary SFP validation dataset (see detailed description in "Materials and Methods" under Maize Sequence Validation Dataset"). The secondary validation dataset of shared probes contains 8204 non-SNP probes and 835 SNP probes (9039 total probes). Of the 20 666 possible pairwise probe comparisons, there is potential to detect 1409 SFPs. Analysis of the shared probes dataset enabled us to compare the SFP detection power of each method without any probeset biases because all of the analyzed validation probesets had signal intensities greater than background on all CF, MF, and mRNA GeneChips. Amplified

Table 2. Mixed model analysis of single-feature polymorphism (SFP) detection power.

FDR†	CF‡		MF ‡		mRNA§		AFLP‡	
	Power¶	No. SFP#	Power	No. SFP	Power	No. SFP	Power	No. SFP
	%		%		%		%	
5%††	1	26	1	20	–	–	45‡‡	17
10%	3	78	14	284	2	29	45	17
20%	30	734	26	514	26	447	45	17
30%	41	1002	37	736	34	573	45	17
40%	49	1179	43	869	39	662	45	17

†FDR, false discovery rate.

‡Array error term *t* test. CF, C_0t filtration; MF, methylation filtration; AFLP, amplified fragment length polymorphism.

§Probeset error term *t* test.

¶SFP detection power was calculated as the proportion of detected true positive SFPs to the total number of potential SFPs at an empirically determined FDR using the primary SFP validation dataset. Details as to how empirical FDRs were determined are provided in the "Materials and Methods" section under "Detecting SFPs in Hybridization Data."

#Number of SFP detected at an empirically determined FDR.

††Power estimates at 5% FDR are less statistically reliable due to the lower number of detected SFPs.

‡‡All power estimates for AFLP were based on a low number of observations and are therefore less statistically reliable than those for the other methods.

Table 3. Mixed model analysis of single-feature polymorphism (SFP) detection power with shared probes.

FDR†	CF‡		MF‡		mRNA§	
	Power¶	No. SFP#	Power	No. SFP	Power	No. SFP
	%		%		%	
5%††	–	–	1	21	–	–
10%††	3	36	4	53	1	9
20%	34	475	23	330	24	337
30%	42	598	35	498	33	462
40%	48	680	43	603	38	536

†FDR, false discovery rate.

‡Array error term *t* test. CF, C_0t filtration; MF, methylation filtration.

§Probeset error term *t* test.

¶SFP detection power was calculated as the proportion of detected true positive SFPs to the total number of potential SFPs at an empirically determined FDR using the secondary SFP validation dataset. Details as to how empirical FDRs were determined are provided in the "Materials and Methods" section under "Detecting SFPs in Hybridization Data."

#Number of SFP detected at an empirically determined FDR.

††Power estimates at 5 and 10% FDRs are less statistically reliable due to the lower number of detected SFPs.

fragment length polymorphism was not included in the shared probes analysis due to the low number of validation probes shared with the other three methods. The results of the shared probes analysis (Table 3) are similar to those of the initial complete datasets (Table 2), with the exception that a reduction in probe numbers eliminated SFP detection power at 5% FDR for CF. In addition, based on results presented in Tables 2 and 3, SFP detection power is reduced 10% at 10% FDR for MF in the shared probeset analysis. These observed losses of power are mainly due to the removal of probes from the complete validation dataset that detected true positive SFPs (5 to 10% FDR) on CF and/or MF GeneChips.

SNP Position Effect

Results of SFP detection power reported in Table 2 indicate that with any one of the target preparation methods, a large proportion (51–61%) of potential SFPs resulting from SNPs remains undetected. The location of a SNP within the 25-bp probe affects target binding efficiency and in so doing also affects PM probe signal intensity. Single nucleotide polymorphism position is defined as the position from the edge of the probe. Position 1 is the first base at either end, and position 13 is the center of the probe. Single nucleotide polymorphisms within the internal 15 bases (positions 6–20) have been found to reduce hybridization much more than nucleotide mismatches within the external 5 bases (positions 1–5 and 21–25) (Kirst et al., 2006; Ronald et al., 2005; Rostoks et al., 2005).

We investigated the impact of SNP position on SFP detection for 984 probes that recognize only a single SNP on hybridizing to the B73, Mo17, and/or CML69 target sequence on CF, MF, and mRNA GeneChips. Of the 984 probes in the probeset dataset, 38% (376) and 62% (608) detect an edge SNP and internal SNP, respectively. The percentage of detected and undetected SFPs resulting from either edge or internal SNPs was calculated (Table 4). Detected SFPs (78–85%) are primarily the result of internal SNPs, whereas undetected SFPs represent an approximate 1:1 ratio of edge-to-internal SNPs. Thus, as expected, the data summarized in Table 4 show that SFPs are called more often if the SNP occurs in the internal region. Also, the percentage of detected SFPs resulting from an edge SNP increases as FDR approaches 40%. Single nucleotide polymorphism position effects are similar for CF, MF, and mRNA. We also examined whether probes detecting multiple SNPs (2, 3, or 4 SNPs) are detected at the same rates as probes detecting a single SNP. Based on analyzed SFP data, the former are called as SFPs no more or less frequently than the latter (data not shown).

Detection Rate of Polymorphic Probesets

We also examined whether it is more effective to identify probesets (probeset-level analysis) that contain one or more polymorphic probes rather than individual SFPs (probe-level analysis). One rationale for this analysis is that as the number of polymorphic probes in a probeset increases, so does the difficulty of identifying specific SFPs. This difficulty

Table 4. The distribution of single nucleotide polymorphism (SNP) position in probes that detect a single SNP.

FDR†	Position‡	CF§		MF§		mRNA	
		No. probes	%	No. probes	%	No. probes	%
5%	Edge	0	0	1	8	0	0
	Internal	0	0	12	92	0	0
10%	Edge	0	0	4	17	0	0
	Internal	16	100	19	83	2	100
20%	Edge	61	19	30	16	23	10
	Internal	258	81	152	84	203	90
30%	Edge	22	24	32	24	16	18
	Internal	69	76	101	76	72	82
40%	Edge	19	32	28	37	17	27
	Internal	41	68	47	63	46	73
Total detected¶	Edge	102	21	95	22	56	15
	Internal	384	79	331	78	323	85
Total not detected	Edge	274	55	281	50	320	53
	Internal	224	45	277	50	285	47

†FDR, false discovery rate.

‡Edge, 1- to 5-bp or 21- to 25-bp SNP position within probe; Internal, 6- to 20-bp SNP position within probe.

§CF, C_0t filtration; MF, methylation filtration.

¶Probe position distribution for combined total detected and total not detected dataset: Edge 38% (376) and Internal 62% (608).

stems from the fact that a target binds weakly when not identical to the probe. As a result, polymorphic probes do not provide an unbiased estimate of DNA (CF, MF, and AFLP) or gene expression levels (mRNA). And yet an accurate estimate of DNA or gene expression levels is required to determine which probes are polymorphic. In addition, a single probe comparison between two lines involves six data points, whereas a probeset comparison involves 90 data points. For these reasons, we hypothesized that a probeset analysis would be far more powerful than the analysis of individual probes.

To estimate the power to detect polymorphic probesets for CF, MF, and mRNA, we constructed a validation set of 289 probesets containing 8 to 15 probes with matching Panzea sequence, of which 109 (38%) contained at least one polymorphic probe (see detailed description in "Materials and Methods" under "Maize Sequence Validation Dataset"). Amplified fragment length polymorphism was not included in the probeset level analysis due to the low number of AFLP probesets called Present and shared in common with the other three methods' arrays. The intensity data for probes within these probesets were analyzed using the mixed model. The *p* value from the *F* test of probe by line interaction was recorded for each probeset and used to rank them in ascending order. Power to detect polymorphic probesets for the three target methods was calculated and is summarized in Table 5. Irrespective of target preparation method, in this study Maize GeneChips are more effective in identifying polymorphic probesets than they are in detecting SFPs (Table 5). Compared with mRNA (19–68%), gain in power over SFP detection with CF (35–38%) and MF (22–43%) is not as dramatic because DNA-based preparation methods should result in more normalized target copy number ratios. Even though the impact of poor DNA or gene expression level estimates is minimized when detecting polymorphic probesets, one significant downside is that individual polymorphic probes are not identified as markers.

DISCUSSION

Conventional methods for SNP discovery in large-scale association mapping studies rely on resequencing candidate gene alleles across distinct individuals of a test population, followed by scoring known SNPs on individuals using one of several array-based SNP genotyping technologies (reviewed in Syvänen, 2005). Expression arrays, however, may offer a more rapid and cost-effective approach. Affymetrix GeneChip expression arrays hybridized with total genomic DNA have successfully functioned as both a polymorphism discovery and genotyping system in *Arabidopsis* and yeast (Hazen and Kay, 2003). Here, we tested whether the Affymetrix GeneChip is appropriate for highly parallel genotyping of larger, more complex genomes such as maize. The Maize GeneChip was evaluated as a high-density platform to detect SFPs in cRNA or DNA hybridization data from three diverse maize inbred lines (B73, Mo17, and CML69).

Targets enriched for gene content and/or reduced in genome complexity were generated by MF, CF, mRNA, and AFLP as a means to score SFPs across the retrotransposon-rich maize genome, but only modest SFP detection power was achieved when these targets were hybridized to the Maize GeneChip. For example, only 39% of expected SFPs were scored with cRNA at 40% FDR—far fewer than the previously reported ~70 to 80% of known sequence polymorphisms scored as SFPs using maize or barley cRNA (Cui et al., 2005; Kirst et al., 2006; Rostoks et al., 2005). The extent of GeneChip replication (Kirst et al., 2006; Rostoks et al., 2005), sampling of multiple tissues (Rostoks et al., 2005), and conservative 5 percentile cutoff (Cui et al., 2005) are the major experimental and data analysis demarcations leading to higher sensitivity in these other cRNA-based SFP studies. In the seminal *Arabidopsis* SFP work of Borevitz et al. (2003), at least 57% of known polymorphisms were detected at 13% FDR with labeled total genomic DNA as the target. Of the DNA-

Table 5. Mixed model analysis of polymorphic probeset detection power.

FDR†	CF‡			MF‡			mRNA		
	No. PP§	Power¶	Power gain#	No. PP	Power	Power gain	No. PP	Power	Power gain
		—%			—%			—%	
5%††	–	–	–	–	–	–	21	19	19
10%	40	37	34	39	36	22	76	70	68
20%	72	66	36	56	51	25	82	75	49
30%	84	77	35	72	66	29	88	81	47
40%	95	87	38	94	86	43	90	83	44

†FDR, false discovery rate.

‡CF, C_0t filtration; MF, methylation filtration.

§No. PP, number of detected polymorphic probesets at an empirically determined FDR.

¶Polymorphic probeset detection power was calculated as the proportion of detected true positive SFPs to the total number of potential SFPs at an empirically determined FDR. Details as to how empirical FDRs were determined are provided in the "Materials and Methods" under "Detecting SFPs in Hybridization Data."

#Power gain, the percent gain in detection power, was calculated as polymorphic probeset detection power (%) minus SFP detection power (%) in Table 2 at an empirically determined FDR.

††Power estimates at 5% FDR are less statistically reliable due to the lower number of detected polymorphic probesets.

based methods evaluated here, MF, CF, and AFLP detected anywhere from 26 to 45% of SFPs at 20% FDR.

What factors are responsible for reducing SFP detection power in this study? Sequencing errors in the Panzea database may be one such factor, if such errors reduced overall detection power by generating undetectable false SFPs. Every effort, however, was made to filter out such sequencing errors before assessing power. As noted in previous SFP studies (Kirst et al., 2006; Ronald et al., 2005; Rostoks et al., 2005), we found that SFPs are detected more robustly if a nucleotide polymorphism in a target sequence binds within the internal 15 bases of the complementary PM probe, whereas edge SNPs are less frequently detected below 40% FDR. The actual minimization of power by this SNP position phenomenon was not quantified in the present study. The binding of spurious nontarget repeat DNAs and multigene family member sequences to probes represents another potential source of genotyping error, compromising power and FDR. In addition, increasing the number of GeneChip replicates has been shown to improve power and FDR (Borevitz et al., 2003; Rostoks et al., 2005), and no doubt this study would have benefited from the same.

Despite the modest detection sensitivity when compared with SFP experiments using smaller genome species, this study marks the first report of using genome-filtered DNA targets to reliably identify more than 10 000 SFPs in a plant genome that contains at least 75% LTR retrotransposons (San Miguel et al., 1996) and is 20X the size of *Arabidopsis*. Based on SNP diversity of maize sequences in the primary SFP validation dataset, we determined that 8.2% (2677/32 511) of all pairwise probe comparisons involve a SNP probe (SFP diversity). Using the power results presented in Table 2 and measure of SFP diversity (0.082), we estimated the number of probes from probesets called Present (MAS5) that would be correctly identified as true SFPs on the Maize GeneChip (Table 6). We then analyzed probe intensity data from Present probesets with the mixed model to determine the observed number of SFPs detected on entire GeneChips. The *p* value cutoffs from the primary SFP validation dataset were used to determine the number of detected SFPs at each FDR. The number of observed true SFPs was in turn calculated by multiplying the number of SFP detected by (1 – FDR). The difference between the estimated and observed number of SFPs can be accounted for by the fact that the estimate of SFPs is founded on SNP diversity and does not include insertion–deletion (indel) diversity, whereas observed SFP numbers account for indels. Kirst et al. (2006) reported that indels represent 40% of all polymorphisms occurring between PM probe and maize target gene sequences.

In most cases, a 10% or lower FDR is acceptable when array genotyping individuals for an association study, but this is highly dependent on sample size, marker density, and levels of genomewide LD. For example, MF is estimated to identify more than 6000 SFPs between the three maize inbred lines at 10% FDR, which results in a cost of ~$0.38 per SFP ($2250/9 arrays). After the initial investment to identify SFPs, the cost per SFP dramatically lowers to ~$0.04 because subsequent genotyping requires only one array per individual (Borevitz et al., 2003). These estimated costs per SFP are very competitive to those reported for the ATH1 GeneChip (~$0.30 per SFP and ~$0.05 per SFP) in 2003 by Borevitz and colleagues. At 20% and higher FDRs, CF detects 1.3X more SFP than MF; however, these more liberal error rates are undesirable for most marker applications.

Although AFLP has far greater detection power from 5 to 20% FDRs, the AFLP design tested here has inferior SFP detection potential and thus does not constitute an economical means of scoring SFPs on the Maize GeneChip. Even though the amplified target fraction contains about 5% of the maize genome (125 Mb/2500 Mb), most amplicons are nongenic, random sequences that result in 4% of probesets called Present. On the other hand, CF and MF are highly preferable to labeling total genomic DNA for a large genome plant species (Rostoks et al., 2005; Buckler and Kirst, unpublished data, 2004) and are

Table 6. Estimated and observed number of true single-feature polymorphisms (SFPs) that were identified on the whole Maize GeneChip.

FDR†	CF‡ No. SFP		MF‡ No. SFP		mRNA No. SFP		AFLP‡ No. SFP	
	Estimated§	Observed¶	Estimated	Observed	Estimated	Observed	Estimated#	Observed
5%††	549	1385	478	992	–	–	1072	918
10%	1647	3046	6698	10 248	712	661	1072	1130
20%	16 474	26 646	12 439	18 454	9259	12 702	1072	1056
30%	22 515	35 422	17 701	25 392	12 108	15 982	1072	1448
40%	26 908	40 729	20 572	29 726	13 889	17 054	1072	1493

†FDR, false discovery rate.

‡CF, C_0t filtration; MF, methylation filtration; AFLP, amplified fragment length polymorphism.

§The estimated number of true SFPs detected on the whole array was calculated by multiplying the probability (0.082) that a pairwise probe comparison involves a SNP probe, the power results shown in Table 2, and number of probes from probesets called Present by MAS5 for each target preparation method. The estimated number of true SFPs for the entire array is less than observed because insertions–deletions (indels) and gene copy number differences are not taken into account.

¶The observed number of true SFPs was determined for each target preparation method by analyzing probe intensity data from probesets called Present (MAS5) using the mixed model. The *p* value cutoffs from the primary SFP validation dataset were used to determine the total number of detected SFPs on the entire array at each FDR. The number of observed true SFPs was in turn calculated by multiplying the number of SFP detected by (1 – FDR).

#Estimated number of SFPs for AFLP was based on a low number of observations and is therefore less statistically reliable than those for the other methods.

††The estimated number of detected SFPs at 5% FDR is less statistically reliable based on the SFP detection power results shown in Table 2.

recommended for scoring SFPs when using the Maize GeneChip. Compared with the other two methods, CF and MF not only provide for the highest coverage of array probes but also account for the highest numbers of detected SFPs. The bias toward a specific fraction of expressed genes in maize is far less for MF and CF than for mRNA because 95% of maize exons are unmethylated (Rabinowicz et al., 2003) and CF gene enrichment is independent of methylation and gene expression patterns (Peterson et al., 2002b).

Even when the cRNA or DNA target sequence was identical to the PM probe sequence, we observed instances where the MM probe had higher signal intensity. Possible explanations for this unexpected outcome are as follows. First, the quantity of hybridized target sequence may be low, resulting in a PM probe intensity that is difficult to separate from the overall background noise. Most PM probes ineffective for SFP genotyping with mRNA-derived cRNA are hindered by low gene expression levels. Second, spurious hybridization of sequences with high similarity to the MM probe could have masked the true target signal. Compared with GeneChips hybridized with cRNA, all genomic DNA target fractions presumably have higher amounts of spurious repetitive DNAs diluting the PM signal. Based on a previously published repeat analysis of CF and MF maize genome sequencing data, the total number of repeat sequences in MF and CF libraries was 33% (17 419/52 649) and 14% (10 154/71 492), respectively (Whitelaw et al., 2003). While our CF and MF libraries did not meet the exact specifications of those analyzed in the above study, these findings indicate that residual repetitive DNAs are almost certainly cohybridized to CF and MF arrays. In particular, a higher percentage of array probes hybridized with AFLP samples is clearly not useful for scoring SFPs. This is not an unexpected outcome given that the 125-Mb AFLP target fraction has a low percentage of amplified sequences complementary to probe sequences. Whatever the cause, probe pairs for which the target is known to be an exact match to the PM probe and of those that have a large MM/PM signal ratio are most likely ineffective for detecting sequence polymorphisms.

As shown in Table 5, another point of interest lies in the fact that the power to detect polymorphic probesets was much greater than the power to detect individual probes at comparable false discovery rates. At least two factors contribute to this difference. First and foremost is the large amount of data available to test probe by line interaction in a probeset. All 135 data points from 15 probes on nine arrays can be used, whereas a comparison of two lines at a single probe involves only six data points. This discrepancy, however, cannot explain why the gain in power was much greater for the mRNA method than for the DNA methods. A likely explanation is that differences in gene expression levels interfere with the ability to detect probe by line interaction with the mRNA method but not with the DNA methods. We did not take into account varying DNA and gene expression levels when calculating probe intensity differences between lines because we found that doing so resulted in lower power for all methods, even the mRNA method (data not shown).

While CF is broadly applicable to both plants and animals, it is technically challenging to generate reproducible libraries from multiple diverse genotypes and to optimize the method for high-throughput applications. Methylation filtration, on the other hand, is specific to plants, and the level of gene enrichment is species dependent (Rabinowicz et al., 2005). Gel purification of the unmethylated gene-rich fraction of plant genomes is also not highly amenable to rapid processing, and cytosine methylation differences between genotypes are known to create non-SNP polymorphisms (Cervera et al., 2002). Moreover, residual genome complexity consisting of repetitive DNA in both CF and MF samples is believed to have complicated SFP detection in this study.

As discussed above, the target preparation methods evaluated in this study offered only modest power to detect SFPs with the Maize GeneChip. The effective use of such arrays for genotyping complex plant genomes would require several improvements, including custom array designs with additional replication and tiling of probes and more aggressive reduction of genomic complexity than can be accomplished via standard MF and CF approaches (e.g., MF, followed by HC). Amplified fragment length polymorphism is expected to be a more powerful method in such cases, provided that probes are selected from sequences represented in the AFLP sample used for hybridization. By using an AFLP design similar to whole-genome sampling analysis in humans (Kennedy et al., 2003), it may be possible to selectively SNP genotype amplified gene fragments and promote reduction of genome complexity to the desired level.

Acknowledgments

We thank D. Costich, E. Ersoz, J. Mezey, and M. Hamblin for their insights and critical review of the manuscript, N. Stevens for technical editing of the manuscript, and two anonymous reviewers for helpful comments. The authors thank Daniel Peterson (Mississippi State University) for help with $C_o t$ filtration. The AFLP technology is covered by patents, and patent applications owned by Keygene N.V. AFLP is a registered trademark of Keygene N.V. GeneChip, BioPrime and TRIZOL are registered trademarks of their respective owners. This work was supported in part by U.S. National Science Foundation grant DBI-0321467 and USDA-ARS. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

References

Arumuganathan, K., and E.D. Earle. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 9:208–218.

Bedell, J.A., M.A. Budiman, A. Nunberg, R.W. Citek, D. Robbins, J. Jones, E. Flick, T. Rholfing, J. Fries, K. Bradford, J. McMenamy, M. Smith, H. Holeman, B.A. Roe, G. Wiley, I.F. Korf, P.D. Rabinowicz, N. Lakey, W.R. McCombie, J.A. Jeddeloh, and R.A. Martienssen. 2005. Sorghum genome sequencing by methylation filtration. PLoS Biol. 3:e13.

Bernardi, G. 1971. Chromatography of nucleic acids on hydroxyapatite columns. Methods Enzymol. 21:95–139.

Bolstad, B.M., R.A. Irizarry, M. Astrand, and T.P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193.

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

Britten, R.J., and D.E. Kohne. 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161:529–540.

Cervera, M.T., L. Ruiz-Garcia, and J.M. Martinez-Zapater. 2002. Analysis of DNA methylation in *Arabidopsis thaliana* based on methylation-sensitive AFLP markers. Mol. Genet. Gen. 268:543–552.

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

Fu, H., W. Park, X. Yan, Z. Zheng, B. Shen, and H.K. Dooner. 2001. The highly recombinogenic bz locus lies in an unusually gene-rich region of the maize genome. Proc. Natl. Acad. Sci. USA 98:8903–8908.

Fu, H., Z. Zheng, and H.K. Dooner. 2002. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc. Natl. Acad. Sci. USA 99:1082–1087.

Gaut, B.S., M. Le Thierry d'Ennequin, A.S. Peek, and M.C. Sawkins. 2000. Maize as a model for the evolution of plant nuclear genomes. Proc. Natl. Acad. Sci. USA 97:7008–7015.

Geever, R.F., F.R.H. Katterman, and J.E. Endrizzi. 1989. DNA hybridization analyses of a *Gossypium* allotetraploid and two closely related diploid species. Theor. Appl. Genet. 77:553–559.

Hake, S., and V. Walbot. 1980. The genome of *Zea mays*, its organization and homology to related grasses. Chromosoma 79:251–270.

Hazen, S.P., and S.A. Kay. 2003. Gene arrays are not just for measuring gene expression. Trends Plant Sci. 8:413–416.

Ihaka, R., and R. Gentleman. 1996. R: A language for data analysis and graphics. J. Comput. Graph. Statist. 5:299–314.

Irizarry, R.A., B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264.

Kennedy, G.C., H. Matsuzaki, S. Dong, W.M. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M.S. Phillips, M.T. Boyce-Jacino, S.P. Fodor, and K.W. Jones. 2003. Large-scale genotyping of complex DNA. Nat. Biotechnol. 21:1233–1237.

Kirst, M., R. Caldo, P. Casati, G. Tanimoto, V. Walbot, R.P. Wise, and E.S. Buckler. 2006. Genetic diversity contribution to errors in short oligonucleotide microarray analysis. Plant Biotechnol. J. 4:489–498.

Liu, K., M. Goodman, S. Muse, J.S. Smith, E. Buckler, and J. Doebley. 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128.

Liu, W.M., R. Mei, X. Di, T.B. Ryder, E. Hubbell, S. Dee, T.A. Webster, C.A. Harrington, M.H. Ho, J. Baid, and S.P. Smeekens. 2002. Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics 18:1593–1599.

Palmer, L.E., P.D. Rabinowicz, A.L. O'Shaughnessy, V.S. Balija, L.U. Nascimento, S. Dike, M. de la Bastide, R.A. Martienssen, and

W.R. McCombie. 2003. Maize genome sequencing by methylation filtration. Science 302:2115–2117.
Peterson, D.G., S.R. Schulze, E.B. Sciara, S.A. Lee, J.E. Bowers, A. Nagel, N. Jiang, D.C. Tibbitts, S.R. Wessler, and A.H. Paterson. 2002a. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res. 12:795–807.
Peterson, D.G., S.R. Wessler, and A.H. Paterson. 2002b. Efficient capture of unique sequences from eukaryotic genomes. Trends Genet. 18:547–550.
Rabinowicz, P.D., R. Citek, M.A. Budiman, A. Nunberg, J.A. Bedell, N. Lakey, A.L. O'Shaughnessy, L.U. Nascimento, W.R. McCombie, and R.A. Martienssen. 2005. Differential methylation of genes and repeats in land plants. Genome Res. 15:1431–1440.
Rabinowicz, P.D., L.E. Palmer, B.P. May, M.T. Hemann, S.W. Lowe, W.R. McCombie, and R.A. Martienssen. 2003. Genes and transposons are differentially methylated in plants, but not in mammals. Genome Res. 13:2658–2664.
Rabinowicz, P.D., K. Schutz, N. Dedhia, C. Yordan, L.D. Parnell, L. Stein, W.R. McCombie, and R.A. Martienssen. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. Nat. Genet. 23:305–308.
Raleigh, E.A. 1992. Organization and function of the mcrBC genes of Escherichia coli K-12. Mol. Microbiol. 6:1079–1086.
Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler, IV. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98:11479–11484.
Ronald, J., J.M. Akey, J. Whittle, E.N. Smith, G. Yvert, and L. Kruglyak. 2005. Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. Genome Res. 15:284–291.
Rostoks, N., J.O. Borevitz, P.E. Hedley, J. Russell, S. Mudie, J. Morris, L. Cardle, D.F. Marshall, and R. Waugh. 2005. Single-feature polymorphism discovery in the barley transcriptome. Genome Biol. 6:R54.
Saghai-Maroof, M.A., K.M. Soliman, R.A. Jorgensen, and R.W. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley Mendelian inheritance chromosomal location and population dynamics. Proc. Natl. Acad. Sci. USA 81:8014–8018.
SanMiguel, P., A. Tikhonov, Y.K. Jin, N. Motchoulskaia, D. Zakharov, A. Melakeberhan, P.S. Springer, K.J. Edwards, M. Lee, Z. Avramova, and J.L. Bennetzen. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765–768.
Singer, T., Y. Fan, H.-S. Chang, T. Zhu, S.P. Hazen, and S.P. Briggs. 2006. A high-resolution map of *Arabidopsis* recombinant inbred lines by whole-genome exon array hybridization. PLoS Genetics 2:e144.
Steinmetz, L.M., H. Sinha, D.R. Richards, J.I. Spiegelman, P.J. Oefner, J.H. McCusker, and R.W. Davis. 2002. Dissecting the architecture of a quantitative trait locus in yeast. Nature 416:326–330.
Stupar, R.M., and N.M. Springer. 2006. *Cis*-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F_1 hybrid. Genetics 173:2199–2210.
Sutherland, E., L. Coe, and E.A. Raleigh. 1992. McrBC: A multisubunit GTP-dependent restriction endonuclease. J. Mol. Biol. 225:327–348.
Syvänen, A.C. 2005. Toward genome-wide SNP genotyping. Nat. Genet. 37(Suppl.):S5–S10.
Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, and B.S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA 98:9161–9166.
Tsolaki, A.G., A.E. Hirsh, K. DeRiemer, J.A. Enciso, M.Z. Wong, M. Hannan, Y.O. Goguet de la Salmoniere, K. Aman, M. Kato-Maeda, and P.M. Small. 2004. Functional and evolutionary genomics of Mycobacterium tuberculosis: Insights from genomic deletions in 100 strains. Proc. Natl. Acad. Sci. USA 101:4865–4870.
Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau. 1995. AFLP: A new technique for DNA fingerprinting. Nucleic Acids Res. 23:4407–4414.
Werner, J.D., J.O. Borevitz, N. Warthmann, G.T. Trainer, J.R. Ecker, J. Chory, and D. Weigel. 2005. Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proc. Natl. Acad. Sci. USA 102:2460–2465.
White, S.E., and J.F. Doebley. 1999. The molecular evolution of *terminal ear1*, a regulatory gene in the genus *Zea*. Genetics 153:1455–1462.
Whitelaw, C.A., W.B. Barbazuk, G. Pertea, A.P. Chan, F. Cheung, Y. Lee, L. Zheng, S. van Heeringen, S. Karamycheva, J.L. Bennetzen, P. SanMiguel, N. Lakey, J. Bedell, Y. Yuan, M.A. Budiman, A. Resnick, S. Van Aken, T. Utterback, S. Riedmuller, M. Williams, T. Feldblyum, K. Schubert, R. Beachy, C.M. Fraser, and J. Quackenbush. 2003. Enrichment of gene-coding sequences in maize by genome filtration. Science 302:2118–2120.
Wilcoxon, F. 1945. Individual comparisons by ranking methods. Biometrics 1:80–83.
Winzeler, E.A., D.R. Richards, A.R. Conway, A.L. Goldstein, S. Kalman, M.J. McCullough, J.H. McCusker, D.A. Stevens, L. Wodicka, D.J. Lockhart, and R.W. Davis. 1998. Direct allelic variation scanning of the yeast genome. Science 281:1194–1197.
Wolyn, D.J., J.O. Borevitz, O. Loudet, C. Schwartz, J. Maloof, J.R. Ecker, C.C. Berry, and J. Chory. 2004. Light-response quantitative trait loci identified with composite interval and eXtreme array mapping in *Arabidopsis thaliana*. Genetics 167:907–917.
Yao, H., Q. Zhou, J. Li, H. Smith, M. Yandeau, B.J. Nikolau, and P.S. Schnable. 2002. Molecular characterization of meiotic recombination across the 140-kb multigenic *a1-sh2* interval of maize. Proc. Natl. Acad. Sci. USA 99:6157–6162.
Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2003. High-Cot sequence analysis of the maize genome. Plant J. 34:249–255.
Zhao, W., P. Canaran, R. Jurkuta, T. Fulton, J. Glaubitz, E. Buckler, J. Doebley, B. Gaut, M. Goodman, J. Holland, S. Kresovich, M. McMullen, L. Stein, and D. Ware. 2006. Panzea: A database and resource for molecular and functional diversity in the maize genome. Nucleic Acids Res. 34:D752–D757.
Zhou, Y., T. Bui, L.D. Auckland, and C.G. Williams. 2002. Undermethylated DNA as a source of microsatellites from a conifer genome. Genome 45:91–99.