Maydica 50 (2005): 000-000

DESIGN, PRODUCTION, AND UTILIZATION OF LONG OLIGONUCLEOTIDE MICROARRAYS FOR EXPRESSION ANALYSIS IN MAIZE

J.M. Gardiner[1,*], C.R. Buell[2], R. Elumalai[1], D.W. Galbraith[1,3], D.A. Henderson[3,4], A.L. Iniguez[5], S.M. Kaeppler[5], Jong Joo Kim[4], J. Liu[2], A. Smith[5], L. Zheng[2], V.L. Chandler[1,3]

[1] University of Arizona, Department of Plant Sciences, 303 Forbes Building, Tucson, AZ 85721, USA
[2] The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA
[3] University of Arizona, BIO5 Institute, 411 Forbes Building, Tucson, AZ 85721, USA
[4] University of Arizona, Department of Animal Science, 231 Shantz Building, Tucson, AZ 85721, USA
[5] University of Wisconsin, Department of Agronomy, 1575 Linden Drive, Madison, WI 53706, USA

Received February 17, 2005

ABSTRACT - Analysis of gene expression on a genome scale can provide useful insights into plant growth and development, and an understanding of the mechanisms used by plants to cope with biotic and abiotic stress. To facilitate analysis of genome-wide gene expression in maize, we have assembled a large collection of maize EST and genomic sequences, designed a set of 57,442 maize 70-mer oligonucleotides to represent these sequences, and printed a two-slide microarray set (MOA and MOB) which is available to the maize research community at minimal cost. To monitor array quality, we have developed a series of printing controls and procedures that when coupled with a 9-mer hybridization assay, allow tracking of spot morphology and printing pin carryover. An optimized hybridization protocol has been developed by testing a series of hybridization temperatures and performing detailed statistical analyses. To facilitate management of all long-oligonucleotide associated array data, Zeamage, a Sybase relational database has been developed and is available at www.maizearray.org. Zeamage contains the appropriate tables and fields for tracking the oligonucleotide sequences and associated annotation, array design, and biological information associated with the microarray hybridizations. The www.maizearray.org website provides additional information on the project, array content, and data analysis tools.

KEY WORDS: Maize; Oligonucleotides; Microarrays; Gene expression.

* This work was supported by the National Science Foundation Plant Genome Research Program (Grant Number DBI-0321663)

* For correspondence (Fax: +1 520 621 7186; e.mail: gardiner@ag.arizona.edu).

INTRODUCTION

Maize is one of the most economically important cereal crops and is grown worldwide with cultivars adapted to a wide variety of growing conditions and climates. The U.S., Brazil, China, and the European Union account for over 70% of the maize grown worldwide. In the U.S. alone, 3.8 billion bushels of corn were produced for livestock in 2003, accounting for 57% of corn production with ethanol and high fructose corn syrup accounting for an additional 11% and 5%, respectively (http:www.ncga.com).

Maize is well suited for addressing biological research questions due to its well developed genetic (SHARAPOVA *et al.,* 2002) and physical (CONE *et al.*, 2002) maps, diploid genetics, and ease of cultivation. The maize genome is ~2.3 GB (BENNETT *et al.,* 1995) and like many crop genomes, is thought to have arisen from genome duplication (GAUT and DOEBLEY, 1997). The maize genome contains a large amount of repetitive DNA with estimates of 57.9% (MESSING *et al.,* 2004) to 62.5% (WHITELAW e*t al.*, 2003) of its genome being comprised of long terminal repeat (LTR) retrotransposon families. Current estimates for the genic (protein coding) regions of the maize genome range from 5% (MEYERS *et al.*, 2001) to 7% (MESSING *et al.*, 2004). The remainder of the maize genome is thought to be composed of uncharacterized highly degenerate retrotransposons, novel retrotransposon LTRs, DNA transposons, and lower-copy noncoding sequences of structural importance (MEYERS *et al.,* 2001).

Considerable interest exists in developing tools and technologies for global analysis of gene expression in maize. These measurements can provide the basis not only for understanding mechanisms in

which regulation of transcript abundance controls plant development, and responses of the plant to biotic and abiotic stimuli, but also for the rational design of strategies to improve crop yield and quality. Since their introduction in 1995 (SCHENA *et al.*, 1995), DNA microarrays have provided an increasingly popular means for providing global measurements of transcript abundance. This type of microarray is comprised of DNA elements designated as "probes" (PHIMISTER, 1999) robotically arrayed on to the surface of a solid support, typically glass. Ideally, each DNA element is designed to represent sequence information unique to an individual gene. The arrays are queried, using fluorescent cDNA designated as "target" (PHIMISTER, 1999) produced from an RNA population of interest, via hybridization under conditions of controlled stringency. The intensities of the immobilized targets at the individual array locations provide the raw data from which changes in transcript abundance can be determined, as a function of the experimental conditions from which the target RNA was produced (DEYHOLOS and GALBRAITH, 2002).

In terms of probe design, the earliest microarrays employed double-stranded PCR amplicons produced from EST collections. In a previous National Science Foundation supported project (DBI 9872657), we initiated production and distribution of cDNA-based maize amplicon microarrays using ESTs from several tissue-specific cDNA libraries (FERNANDES *et al.*, 2002). These were ultimately condensed into a single amplicon microarray comprising 25,000 unigenes. The main advantage of the amplicon-based microarray is its simplicity, in that probe design does not require complete genomic sequence information. A secondary advantage is that these arrays are generally tolerant to DNA polymorphisms, making it possible to compare gene expression patterns across genotypes (CASATI *et al.*, 2003; YU *et al.*, 2003). The major limitation of amplicon microarrays is the occurrence of cross hybridization between genes sharing sequence identity. Empirically, genes having more than 70% sequence similarity, or patches of sequence identity exceeding 20 bp, exhibit cross hybridization (XU *et al.*, 2001). A second issue is the presence of complementary probe strands immobilized at each array element location, which can potentially serve to compete with target hybridization, thereby complicating the kinetics of the process. Finally, the hybridization characteristics of different target:probe pairs will inevitably exhibit different melting temperatures, which impede optimization of stringencies.

The increasing availability of whole genome sequence information has allowed microarray probes to be produced from single-stranded oligonucleotides, typically 50-70 bases in length (hereafter "long-oligonucleotide"). Use of single-stranded probes eliminates problems of probe complementarity, and design strategies can be employed that aim both at a unified Tm value and at maximizing the degree to which hybridization specifies single gene transcripts. An additional advantage of the long-oligonucleotide approach is that the clone archiving and PCR steps necessary to produce cDNA arrays are bypassed. Long-oligonucleotides also appear insensitive to single nucleotide polymorphisms, a feature particularly important in maize, which has a high level of polymorphism (BUCKLER *et al.*, 2001). Design of long-oligonucleotide probe sets is conceptually most complete when the entire genome sequence is available (cf. Arabidopsis or rice) and there is accurate annotation of transcribed sequences. Nevertheless, long-oligonucleotide probe sets can also be designed in the absence of complete genome sequence, which is the current situation for maize, using the available sequence information, the majority of which reflects the transcriptionally active portion of the genome.

Here we report the design and fabrication of maize long-oligonucleotide arrays for public sector researchers. We report our approach for DNA sequence selection, assembly, and oligonucleotide design, including experiments with 70-mer oligonucleotide arrays to assess orientation and expression of the genes for which long oligonucleotides were designed. We describe our microarray printing and quality assurance procedures, and experiments to determine the optimal hybridization procedures. Information is also presented on Zeamage, a project-specific relational database which allows data curation, mining, and dissemination to the maize research community via the www.maizearray.org website.

MATERIALS AND METHODS

RNA preparation

Total RNA for both the NimbleGen and spotted long-oligonucleotide experiments was isolated using Trizol (Invitrogen, Carlsbad, CA) from tissue samples pulverized in liquid nitrogen. Poly(A^+) mRNA for spotted long-oligonucleotide experiments was purified using DynaBeads Oligo dT 25 (Dynal Biotech, Oslo, Norway) according to manufacturer's instructions.

Target labeling and hybridization methods

Target labeling and hybridization to NimbleGen arrays was done using standard NimbleGen protocols (Nuwaysir *et al.*, 2002). For the spotted long-oligonucleotide arrays, we employed an indirect target labeling method, in which polyA RNA was converted into first strand cDNA using random hexamer primers with 5'-(3-aminoallyl)-2'deoxyuridine-5'-triphosphate (AA-dUTP) (Ambion, Austin, TX) supplemented into the reaction. PowerScript reverse transcriptase (BD Biosciences Clontech, Palo Alto, CA) was used according to manufacturer's instructions for the first strand cDNA synthesis reaction. After the reaction was completed the unincorporated nucleotides and hexamer primers were removed using a Qiaquick PCR purification column (Qiagen, Valencia, CA). The AA-dUTP that was incorporated into cDNA was then coupled to either Cy3 or Cy5 mono reactive dyes (Amersham Pharmacia, Piscataway, NJ). Dye coupling to cDNA was monitored using a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, DE). A complete description for this and other maize microarray protocols described below is provided at http://www.maizearray.org.

Prior to hybridization, slides were held face down over a 60°C water bath for 10 seconds to re-hydrate the array elements and then snap dried on an 80°C heating block for 3 to 10 seconds. This process was monitored by visual inspection and was repeated at least 4 times or until all the spots were evenly rehydrated. Oligo DNA was cross-linked to the glass slide using a Stratalinker (Stratagene, La Jolla, CA) by exposing the microarray slides (DNA side up) under a UV light, with the energy setting at 180mJ/cm^2 of 254-nm UV-C radiation. Slides were washed with 1% (w/v) SDS for 5 min at room temperature using an orbital shaker, followed by multiple rinses with nanopure water and a final 100% ethanol rinse. The slides were dried by centrifugation at 200X G for 2 to 5 min and used immediately for hybridization. Hybridization mix (250 µl) was prepared by mixing 25 µl 20X SSC, 15 µl Liquid Blocking Reagent (Amersham, Piscataway, NJ) 10 µl 2% SDS, 160 µl Cy5 and Cy3 labeled targets, and 40 µl H_2O in a 1.5mL tube. Prior to hybridization, cDNA targets were denatured by heating the hybridization mix at 94°C for 2 minutes and then stored on ice until hybridization.

Microarray hybridizations to determine optimal hybridization temperature were performed at 50, 55, and 60°C for 12 hours. After hybridization, microarray slides were washed for five minutes with 2XSSC, 0.1% SDS, at 55°C for the 55 and 60°C hybridizations, and 50°C for the 50°C hybridization. The second and third washings for all three temperature treatments were carried out at room temperature with 0.5XSSC and 0.1X SSC respectively. Microarray slides were scanned immediately after washing, using an Axon 4100AL scanner (Axon Instruments, Union City, CA) with the same settings of laser power and photomultiplier tube gain for all slides. Median signal intensities were calculated for each spot using the GenePix 6.0 software (Axon Instuments, Union City, CA). Spot values were measured and raw signal intensity was adjusted by subtracting local background. In some instances, further background adjustment of the raw signal intensity was accomplished by subtracting 100 to 1000 in 100 unit increments. Spots were declared positive if the signal intensity values were higher than the sum of background and correction thresholds.

For random 9-mer hybridizations to determine overall slide quality and spot morphology, slides were hybridized with a preheated (70°C for 2 min) cy3-9mer-hybridization mix (12 µl 20 X SSC, 5.0 µl Amersham Liquid Block, 2 µl 2% SDS, 20 µl Cy3-labeled random 9mer (Qiagen; Valencia, CA), and 61 µl H_2O) using a lifter slip (Erie Scientific, Portsmouth, NH). Hybridizations were done in the dark for 1 to 2 hours at room temperature. Slides were washed with washing solution 1 (2 X SSC, 0.1% SDS) for 5 minutes at room temperature, followed by washing solution 2 (0.5 X SSC) for 5 minutes at room temperature. Slides were dried and scanned as described above.

Applied statistical models

A Gaussian mixed linear model was applied to test for differences in number of positive spots between temperatures. Response was recorded as either 1 or 0, depending on whether the adjusted values after signal and background correction were above 0 or not. The mixed model used to estimate proportions and variances was: $y_{ijk} = \mu + T_i + A(T)_{ij} + e_{ijk}$, where y_{ijk} is adjusted signal value on spot k of slide j under temperature i, µ is overall mean across temperatures, arrays, and genes, and T_i is the effect of i^{th} temperature. $A(T)_{ij}$ is jth array effect nested within i^{th} temperature, and e_{ijk} is a residual. In the model, temperature effect was treated as fixed and the array effect within temperature as random. This allowed for the estimation of array variance at each hybridization temperature.

RESULTS AND DISCUSSION

Microarray Element Design

DNA sequences used for long-oligonucleotide design

To maximize the identification of the largest number of genes included on the array in the absence of complete genome sequence, multiple sources of maize sequence were used to design the long-oligonucleotides. In addition to a collection of Expressed Sequenced Tags (ESTs)/Expressed Transcripts (ETs), a large collection of "gene enriched" genomic sequences were available as a result of a genomic sequencing project that was focused on sequencing the gene rich portions of the maize genome (Whitelaw *et al.*, 2003). These sequences were clustered to become the TIGR Assembled Zea Mays (AZMs) and were of particular interest because they were likely to contain genes not represented in the EST assemblies. Additional maize sequences available for long-oligonucleotide design included repeat, chloroplast, and mitochondrial sequences.

The primary source of sequences for design was the TIGR Maize Gene Index Release 13.0 (ZmGI 13; Quackenbush *et al.,* 2001). This release contained a total of 55,063 sequences; 27,607 tentative consensus sequences (TCs, assembled clusters) and 27,456 singleton ESTs/ETs. One requirement associated with selecting sequences for oligonucleotide design is that the correct strand be selected. The TIGR Gene

Indices are oriented into coding or non-coding strands based on multiple data types such as 5' and 3' information in the Genbank record, presence of a polyA or T tail, and alignment with protein sequences in a non-redundant amino acid database. For ZmGI 13, only 29,286 (53.2%) of the 55,063 total unique sequences were oriented; 20,236 (73%) of the TCs were oriented and 9,050 (32.9%) of singleton ESTs/ETs were oriented. Investigation of this low representation of orientation information revealed that a majority of the maize ESTs deposited into the dbEST division of Genbank did not contain orientation information, thereby greatly reducing the data available to orient the sequence. As a consequence, we employed additional computational approaches to predict the orientation of the ZmGI 13 sequences. We searched the sequences against the predicted Arabidopsis proteome (www.tigr.org, Release 4.0), the predicted rice proteome (rice.tigr.org, Release 1), all oriented monocot TIGR gene index sequences (barley, onion, rice, rye, sorghum, wheat), and the open reading frames predicted from the TIGR AZM sequences (WHITELAW *et al.* 2003; Release 3). A "weighted voting scheme" was employed to deduce the orientation of the ZmGI 13 sequences. From this effort, another 15,301 sequences could be oriented. Thus, in total 44,587 (81%) sequences were oriented from ZmGI 13, leaving 10,476 (19%) with unknown orientation. A total of 4,127 of these unoriented sequences were subsequently oriented using experimental evidence from the NimbleGen 70-mer oligonucleotide arrays (details provided below) resulting in a total of 48,714 definitively oriented sequences from ZmGI 13. Multiple sequences were present in this set with similarity to sequences found in the maize mitochondrial genome, the plastid genome, and the TIGR Maize Repeat Database 3.0 (http://www.tigr.org/tdb/tgi/maize/repeat_db.shtml). Both organellar and repetitive sequences were removed, resulting in a total of 46,000 sequences from ZmGI 13 used in the Array Design Dataset (Table 1).

Additional long-oligonucleotides were designed from a set of ESTs derived from root cDNA libraries (BOHNERT H., unpublished) that had not been deposited in Genbank when ZmGI 13.was assembled. These ESTs (15,185 total) were assembled using the TIGR Gene Index clustering software (PERTEA *et al.,* 2003) resulting in 7,765 unique sequences. Among them, 1,250 sequences which were not present in the ZmGI 13 dataset (cutoff criteria of 94% identity and 30% of the coverage of TCs/singletons from ZmGI 13) were added to the Array Design Set (Table 1).

TABLE 1 - *Sources of maize sequences from which oligonucleotides were designed for Release 1 of the NSF Maize Oligonucleotide Array Project.*

Source	Number of Sequences
ZmGI Release 13[a]	46,000
Root-derived ESTs[b]	1,250
AZM Release 3[c]	2,000
AZM Release 4[c]	6,731
Organelles	359
Repetitive Sequencesd	804
Pet	297
Transgenes	11

[a] The TIGR ZmGI Release 13 was filtered for repetitive sequences and organellar sequences. Only those sequences which could be oriented were included in the final oligonucleotide design and array production.

[b] Root-derived ESTs from H. Bohnert were clustered and assembled and represent Genbank accession numbers CF623104-CF6382

[c] TIGR AZMs were obtained from http://www.tigr.org/tdb/tgi/maize/.

[d] Repetitive sequences (402) from the TIGR maize repeat database (http://www.tigr.org/tdb/tgi/maize/repeat_db.shtml) were used to oligonucleotides to both strands.

Another source of maize sequences for oligonucleotide design was provided by maize genomic sequences with predicted genes. Using the TIGR AZMs (Release 3, http://www.tigr.org/tdb/tgi/maize/), we identified genes using the *ab initio* gene finder FGENESH (SALAMOV and SOLOVEY, 2000). The genes were searched against the TIGR Maize Repeat Database and the TIGR ZmGI 13. The 2,000 longest genes that lacked similarity to repetitive sequences or sequences in ZmGI 13 were added to the Array Design Set. During the entire design process, an updated set of AZMs was made available, AZM Release 4. As with the AZM Release 3, we identified putative genes that were not related to repetitive sequences, or present in our existing Array Design Set. As we had an excess of genes because of limited space available on the array, we applied two additional criteria to these genes. Genes from AZM Release 4 had to align with an EST from a monocot species and had to have experimental expression as determined by hybridization on the NimbleGen arrays (details provided below). From these criteria, 6,731 genes were added to the Array Design Set (Table 1).

To provide a complete representation of organellar genes, we collected chloroplast genes ac-

cording to gene annotations for the maize chloroplast genome in Genbank (Accession X86563). We used FGENESH to predict genes in the maize mitochondrial genome sequence obtained from Washington University (ftp://genome.wustl.edu/pub/seqmgr/maize_mitochondria/Z_MTNB.2Jan2003.con). These genes were added to the Array Design Set. To provide a robust, but not comprehensive, set of oligonucleotides for repetitive sequences, we selected 402 repetitive sequences for the Array Design Set that provide a broad representation of the repetitive sequences within the maize genome (Table 1). Long oligonucleotides were designed in both orientations for each of these repeat sequences. These repetitive sequences on the array can be assigned to the following classes: 193 class 1 (retrotransposons), 69 class 2 (DNA type transposons), 29 centromere associated sequences, 13 rRNA, 11 knob associated sequences, 23 others (plastid, genes, telomere, dispersed), and 64 unknown sequences. This repeat set represents a large portion of the known repetitive sequences in maize, but is biased toward high copy and expressed elements. This bias is caused by the type of data used to create the probe set such as BAC sequences (high copy sequences), TIGR maize gene index (expressed sequences), TIGR maize repeat database (many high copy sequences).

We also solicited the maize community for sequences that they would wish to see on the array. These sequences were termed "pet" genes and these were added to the Array Design Set if they were not highly similar to existing sequences in the set. The final Array Design Set contained 57,441 maize sequences (Table 1).

Orientation of unoriented gene index sequences and validation of AZM expression using NimbleGen long-oligonucleotide arrays

We employed experimental evidence generated by NimbleGen arrays to orient the ZmGI 13 sequences of unknown orientation and to validate AZM expression. The NimbleGen maskless array system offers the advantage that small numbers of arrays can be rapidly fabricated which allows for rapid data acquisition (Singh-Gasson *et al.*, 1999), an important consideration in orienting these unoriented ESTs. Oligonucleotide 70-mers were designed in both orientations (designated as orientation 1 and orientation 2) for 9,346 of the unoriented gene index sequences and predicted genes for 8,103 AZMs utilizing the Qiagen/Operon design criteria (described below). Ideally, one of these two orientations would detect significant expression in at least one of the maize tissues assayed. To capture a diverse expression profile, RNA from seedling leaves, roots, 11 DAP endosperm, Black Mexican Sweet suspension culture cells, and HiII callus were used to interrogate twenty NimbleGen arrays for gene expression. After logarithmic transformation of raw intensity values, an analysis of variance (ANOVA) model was used to determine significance of expression as well as orientation. The ANOVA model accounted for the representation of each maize sequence by two probes, one in each orientation, with four technical replicates for each of the five tissues. Therefore, each pair of oligonucleotides (orientation 1, orientation 2) had a total of 40 observations. Expression levels were considered significant if the background corrected mean expression was significantly greater than zero ($p<0.05$). A liberal p value (0.05 as opposed to 0.01) was deliberately chosen to maximize the number of sequences that could be oriented.

The ANOVA revealed that 30% of the 9,346 gene index sequences with unknown orientation did not produce significant signal in any tissue tested (Fig. 1). The same analysis determined that 44% (4127) of the sequences were significantly expressed in only one orientation for one or more tissues, with 22% indicating expression for orientation 1 and 22% indicating expression for orientation 2. These 4,127 definitively oriented sequences were added to the oligo design set. Interestingly, 25% of the sequences showed significant expression for both orientations in one or more tissues. Although sense and antisense expression could not be distinguished in these experiments, this result suggests that a quarter of the unoriented sequences show significant levels of antisense expression. This phenomenon has been previously reported in higher plant systems such as *Arabidopsis thaliana* (Yamada *et al.*, 2003) and rice (Osato *et al.*, 2004) and remains to be further corroborated in maize. About 1% of the oligos showed differential orientation across tissues, suggesting that the orientation expressed for these sequences is tissue dependent.

The majority (82%) of the AZMs on the NimbleGen arrays has significant expression in at least one of the five tissues assayed for the sense orientation predicted by the FGENESH genes (Table 2). This finding provided support for the predicted genes as does the fact that only 11% of the AZMs oligonucleotides had antisense only expression. It is interesting that 54% of the AZMs had significant expression

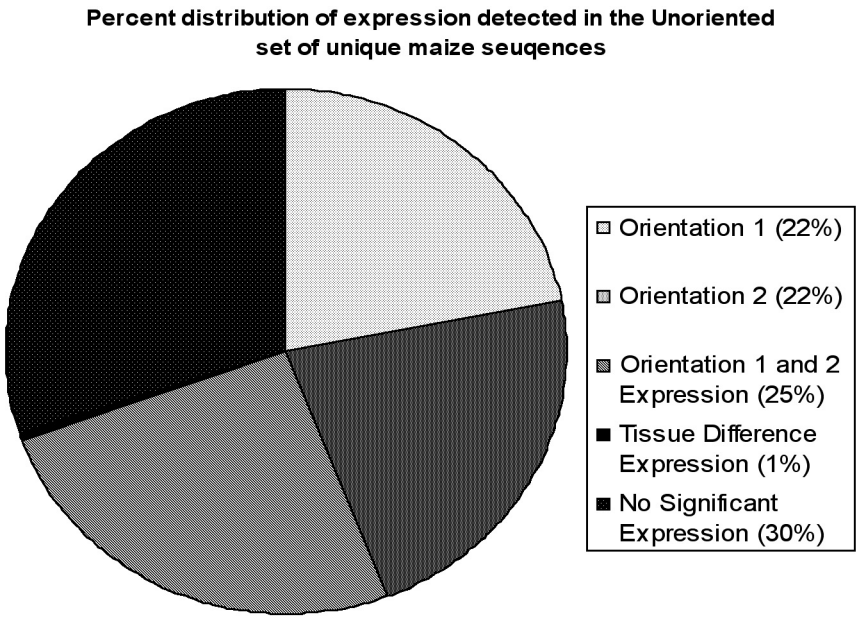

FIGURE 1 - Expression analysis of 9346 unoriented maize gene index sequences for the Maize Oligonucleotide Array project. Each portion of this figure represents the percent of expression and orientation from the 9,346 unoriented sequences. RNA from five maize tissues was used to interrogate high density NimbleGen arrays containing 70-mer oligonucleotides in both orientations for each of the 9,346 unoriented sequences.

TABLE 2 - *Validation of AZM expression*

Group	AZM4s (%)
Sense Expression Only	28
Antisense Expression Only	11
Sense and Antisense Expression	54
No Significant Expression	6
Tissue Specific Orientation Expression	1

To validate expression of 8,103 predicted genes for the AZM genomic sequences, NimbleGen arrays containing long-oligonucleotides in both the FGENESH predicted sense and antisense orientations were interrogated with RNA from five maize tissues. The majority of long-oligonucleotides (82%) detected significant expression in the predicted orientation giving support to the FGENESH predicted genes. A large number of the long-oligonucleotides (54%) detected expression in both orientations.

for both orientations which is higher than the 25% observed for the unoriented ESTs. Due to space limitations on the array, only sense oligos (6,731) indicating expression were added to the array design set.

Oligonucleotide design and selection

Oligonucleotides were designed from the Array Design Set by Qiagen/Operon using a proprietary set of algorithms that identify the optimal oligonucleotide sequence based on an oligonucleotide length of 70, a melting temperature (Tm) of 78°C + 5°C in 0.1 M NaCl, location of the oligonucleotide within 1000 bp of the 3' end of the sequence, cross hybridization with other sequences in the Array Design Set of less than or equal to 70%, sharing of less than 20 identical bases with another sequence, poly N tracts less than 9 bases, and hairpin stem length less than 9 bases. A majority of the oligonucleotides meet all these design criteria, with a limited number failing one or more of these criteria. Qiagen/Operon binned the oligonucleotides into five possible design phases (Phase1, Phase1a, Phase2, Phase2a, Phase3) based on which criteria, or set of criteria, the oligonucleotide met. Phase 1 oligonucleotides (73%, 41,953) met all criteria; Phase 1a oligonucleotides (3%, 1,969) met all criteria except that the oligonucleotide length was 40 or 50 nucleotides and/or Tm range was wider than 78 ± 5°C. Phase 2 oligonucleotides (1.7%, 960) met all criteria except that the position of the oligonucletide was not less than or equal to 1000 bases from the 3' end of the sequence. Phase 2a oligonucleotides (<1%, 126) met all criteria for Phase 2 oligonucleotides except that the oligonucleotide length was 40 or 50 nucleotides and/or Tm range was wider than 78 ± 5°C. Phase 3 (22%, 12,444) was used to describe all other oligonucleotides such as those with sequence similarity greater than 70% to another oligonucleotide within the Array Design Set. Clearly, the final set contains oligonucleotides that will cross-hybridize with non-target sequences, such as those oligonucleotides that contain greater than 70% homology to another oligonucleotide in the design set. A total of 57,452 oligonucleotides were selected for synthesis.

Array Production And Quality Control

Array printing

The production of robotically spotted microarrays requires access to reliable materials and equipment as well as a series of well developed protocols that allow monitoring of array quality. With the current printing technology, using our OmniGrid 300 (Genomic Solutions, Ann Arbor, MI), it is possible to spot ~ 34,000 – 36,000 spots per slide without compromising spot quality. Therefore, we spot the ~57,400 maize oligos onto two microscope slides designated as Maize Oligoarray A (MOA) and Maize Oligoarray B (MOB), at room temperature using 48 SMP3 pins (Telechem International, Sunnyvale, CA) in an environment with 32% relative humidity. The array-ready 5' amino modified long-oligonucleotides were received as aliquots of 300 pm (Operon Biotechnologies) in 384 well plates and resuspended in 15 mL of printing buffer (3X SSC). The OmniGrid 300 can produce 308 slides in a single print run and can print ~1000 slides from a single 300 pm aliquot of oligos. The long oligo microarrays were spotted on either specially coated Powermatrix slides (Full Moon Bisosystems, Sunnyvale, CA) or Superamine slides (Telechem International).

Array quality control

The major challenges in spotted microarray production are avoiding DNA carryover between samples and maintaining uniform spot morphology across the microarray. It is important to clean the spotting pins thoroughly between the samples and dry the pins completely to avoid DNA carryover contamination. Microarray spotting pins are cleaned after every sample using a procedure involving two cycles of washing. A single wash cycle consists of sonication for two seconds, and drying for two seconds followed by 6 cycles of washing and drying, culminating in a final 40 second drying cycle.

To monitor the cleaning process we have developed an assayable system utilizing a Printing Validation plate (PV plate). The PV plate is made up of wells containing12 maize positive control oligos, and wells containing only 3X SSC printing buffer.

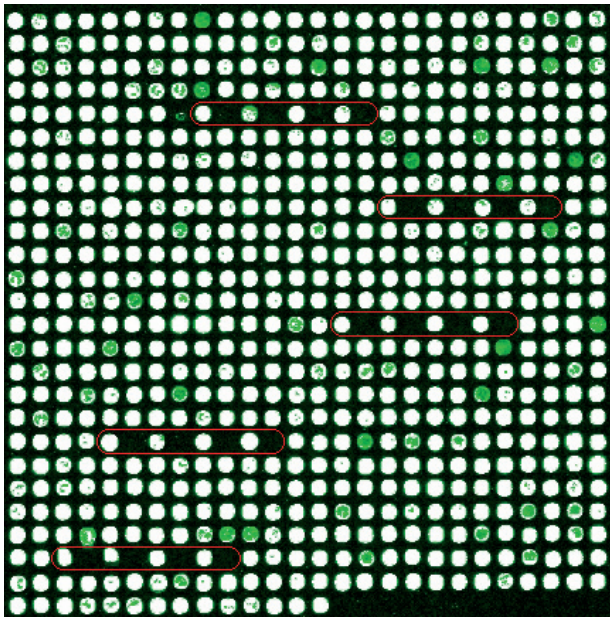

FIGURE 2A - Cy3 Random 9-mer Quality Control Hybridization
To visualize the spotted long-oligonucleotides, and thereby assess print tip cleaning and spot morphology, the microarrays are hybridized using Cy3 labeled 9-mers. A print validation plate with alternating wells containing either long-oligonucleotide DNA or 2X SSC printing buffer is used five times during array printing. A single 26 X 26 subarray is shown and the eight printing validation spots resulting from each insertion of the printing validation plate are outlined in red. The alternating pattern of Cy3 9-mer hybridized spots corresponds to wells with and without long-oligonucleotide DNA, and allows monitoring of print tip cleaning.

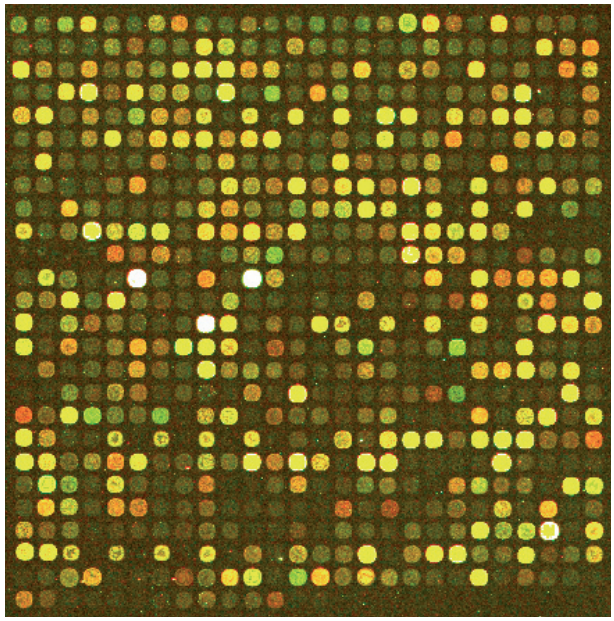

FIGURE 2B - A Long-Oligonucleotide Microarray Hybridized with Cy3 and Cy5 labeled cDNA
A 26 X 26 long-oligonucleotide subarray hybridized with Cy3 labeled (green) leaf cDNA and Cy5 labeled (red) endosperm cDNA. Red and green spots represent the transcripts of genes which are expressed only in endosperm or leaf, respectively. White spots correspond to transcripts of highly expressed genes, whose Cy3/5 signals were saturated.

TABLE 3 - *Quality control analysis of maize oligo arrays.*

Microarray Printing	Good Spots (%)	Bad Spots (%)	Buffer Only Spot Background (pixels)	Average Spot Diameter (µm±S.D.)
MOA-1-1	98.9	1.1	38	116±7.6
MOB-1-1	99.2	0.8	23	117±9.1
MOA-1-2	99.5	0.5	69	102±7.5
MOB-1-2	99.5	0.5	69	102±7.5
MOA-1-3	99.5	0.5	84	89±10.1
MOB-1-3	99.7	0.3	79	97±10.6

To assess the quality of maize long-oligonucleotide arrays printed on the OmniGrid 300, selected arrays from each printing are hybridized with Cy3 labeled random 9-mers. This allows monitoring of print pin cleaning and spot morphology ensuring that high quality long-oligonucleotide arrays are supplied to array users.

The wells containing positive control oligos and printing buffer alone are arranged in the PV plate to produce an alternating pattern of oligo and printing buffer spots; this enables immediate detection of DNA carryover between sequential sample pickups. The PV plate is used at regular intervals (approximately every fifteenth plate) throughout the microarray printing process to comprehensively monitor the effectiveness of pin cleaning (Fig. 2).

To directly assay the effectiveness of the printing process, as well as to monitor DNA carryover, we developed a Cy-3 labeled random 9-mer hybridization protocol (see Materials and Methods) that allows estimation of spot quality and size. In every print run, two slides are selected for random 9-mer staining and hybridization images are generated to estimate the parameters described above (Table 3). Our Quality Control analysis indicates that the maize oligo arrays contained between 98.9 and 99.7% good spots with average spot sizes that ranged from 97 µm to 117 µm in diameter. To evaluate potential DNA carryover, we estimate the average signal for all the buffer-only negative control spots from random 9-mer stained Quality Control slides. If the buffer-only average signal is two-fold greater than the average background signal then we conclude the existence of DNA carry-over caused by insufficient pin cleaning. Our evaluation of DNA carryover thus far has indicated that this has not been an issue in the production of our maize long-oligo arrays. Presumably this is due to the stringent pin washing conditions used in the production of our arrays.

Each maize long oligo microarray slide (MOA and MOB) also contains twenty-two nonhomologous negative controls. Twelve oligos are derived from human genes and ten oligos were obtained from the commercially available SpotReport Alien Oligo system (Stratagene, La Jolla, CA). These negative controls can be used to estimate the non-specific background hybridization or as spiking controls in the target labeling process to allow estimation of bias in dye incorporation and/or allow comparison of results between and within microarray experiments.

Determination of the optimal array hybridization temperature

A well tested, hybridization protocol is essential for obtaining both reproducible and meaningful microarray data. Of particular importance in the hybridization protocol is the combination of temperature and salt concentration which determines the effective Tm of the probe and target. While it is not realistic to assume that any single salt and temperature combination can produce optimal results for all oligos on the array, it is possible to select a set of hybridization conditions that give consistent results with minimal cross-hybridization. To select the optimal hybridization temperature for our 2X SSC hybridization buffer, we tested hybridization temperatures of 60, 55, and 50°C. A large pool of Cy5 labeled cDNA from 14 day old seedlings was generated and hybridized to nine identical arrays, three at each of the three hybridization temperatures.

A detailed statistical analysis was performed on the series of replicated hybridization experiments done at the three temperatures to determine the optimal hybridization temperature for the maize long oligo arrays using a 2X SSC buffer. Both binomial fixed effects (logistic regression) and normal mixed

effects statistical models were used on the binary response variable of spot presence, *i.e.* detectable above background plus some threshold value as described later. Expression values were derived from a range of background threshold adjustments to test for differences in the number of detectable spots (those that are positively hybridizing) and array variance in number of detectable spots at different hybridization temperatures. The binomial fixed effects model supported 55°C as the best hybridization temperature across a range of background adjusted values. The mixed effects model supported 55°C as generally giving the greatest number of detectable spots in many but not all background thresholds. In one set of background adjusted values, the 50°C hybridization gave slightly more detectable spots than 55°C but the variance between arrays was generally larger across all background adjustments for the 50°C degree hybridizations (Fig. 3). This may be due to a greater tendency for cross hybridization at the 50°C hybridization temperature, resulting in the larger observed array variance in detectable spots. Taken in aggregate, a 55°C degree hybridization temperature appears optimal for hybridization as it results in smaller array variance for number of detectable spots across a range of background adjustment thresholds and also produces a large number of detectable spots. Additional support for 55°C as the optimal hybridization temperature is provided by the calculation of a Tm of 63°C in the salt conditions provided by 2X SSC and the generally accepted idea that oligo hybridizations should be performed at 5 to 10°C below the calculated Tm.

Zeamage Database and Bioinformatics Resources

Zeamage database

We developed a Sybase relational database (Zeamage for *Zea mays* microarray gene expression) to manage all data associated with the fabrication and use of maize oligonucleotide arrays. Appropriate tables and fields were created to allow for tracking of oligonucleotide sequence and annotation, array design, biological information associated with hybridizations, array intensity data (raw and normalized), and normalization methods. User interfaces are being developed which will allow array users to deposit all data associated with their hybridizations to the Zeamage database. This will include the biological information pertaining to the study, hybridization conditions, scanning information, and the actual intensity data. As a consequence, the data will be compliant with the Minimum Information About a Microarray Experiment (MIAME) guidelines (Brazma *et al.,* 2001). The intensity data will be processed using LOWESS normalization (Yang *et al.,* 2002) to provide a standard set of hybridization within the database. All data will be made available to the public within a maximum of six months from deposition. Search, browse, and analysis tools are currently being developed so that users can search the database at the study, hybridization, or gene level.

Additional bioinformatics resources

The project website (www.maizearray.org) provides useful information on the project, the arrays, and tools available for the community. Pertinent information on oligonucleotide design and annotation is available through the project web pages. We pro-

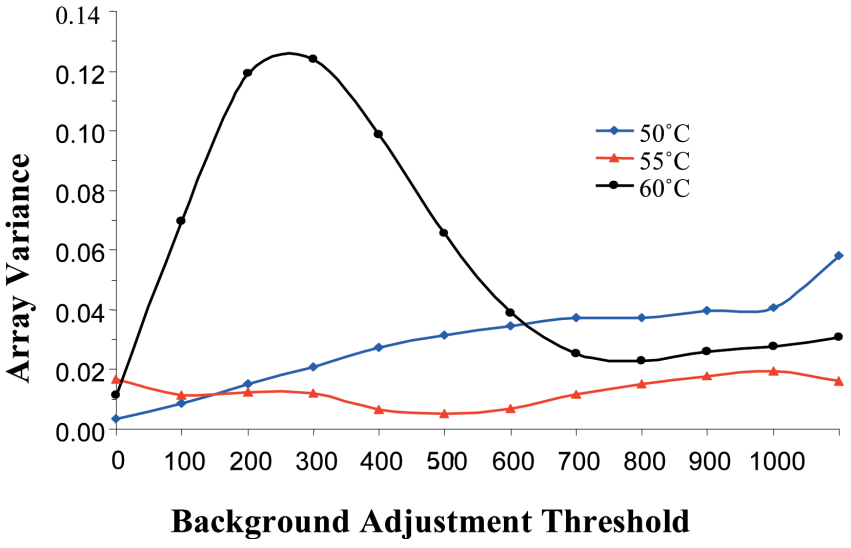

FIGURE 3 - A range if background adjustment thresholds were used to calculate array variance for the number of positively hybridizing spots. The 50 and 55°C hybridizations gave nearly identical numbers of positive spots but the array variance was consistently lower for the 55°C hybridizations suggesting a lower tendency towards cross-hybridization at this temperature.

vide a set of search tools (Basic, Advanced, Batch and BLAST) for the user to query the database for oligonucleotides of interest. A critical feature of the oligonucleotides is robust and current annotation. Clearly, the maize genome is represented by a large set of ESTs and partial genome sequences which are fragmentary and incomplete. Thus, providing manually curated, high quality annotation at this time is not warranted. We have devised an automated method to provide up-to-date annotation with high information content for the oligonucleotide set. Through the project website, we provide a series of annotation data types for the oligonucleotides. These include unique identifiers, Qiagen/Operon identifiers, Tm, design phase, length of oligonucleotide, alignment to maize sequences in the ZmGI, AZM, Maize Repeat Database, Organelles, or "Pet" collection, and putative annotation for that sequence. Other annotation data types include mapping the oligonucleotides to the TIGR Rice Genome Pseudomolecules (rice.tigr.org), mapping to the maize chromosomes, and assigning putative Enzyme Commission numbers. One caveat of the annotation is that the oligonucleotide can align with more than one maize accession; if an oligonucleotide sequence matches multiple accessions at 100% identity and 100% length, then we provide all of those accessions for the user. However, we assign putative annotation to the oligonucleotide only using the longest 100%/100% accession. The top three matches at 95-100% identity and 100% length are also provided along with their putative annotation. The first annotation of the oligonucleotide set was generated immediately after synthesis (Table 4). The annotation is further complicated by continual updating of available maize sequences and the desire to provide the user with the most current annotation. Thus, we "remap" the oligonucleotides to the new releases of the ZmGI and AZMs when sufficient new data is available for this purpose. As a consequence, we have different versions of the annotation (Table 4). The initial annotation (version 1) was employed during the design process described above. Our current annotation, version 2, contains mapping of the oligonucleotides to a new release of the ZmGI (Release 14), release 4 of the AZMs, and release 2 of the TIGR Rice Pseudomolecules.

TABLE 4 - *Annotation and mapping of oligonucleotide sequences to public maize sequences.*

Source	No. Oligonucleotides Version 1	Version 2
ZmGI Release 13/14[a]	46,074	46,723
AZM Release 4[b]	10,074	9,993
Organelles	289	289
Repetitive Sequences	261	292
Transgene	11	11
Others	743	144

[a] The TIGR ZmGI Release 13 (Version 1) and 14 (Version 2) were obtained at http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=maize.

[b] TIGR AZMs obtained from http://www.tigr.org/tdb/tgi/maize/

An additional set of tools and web pages developed for users is the Oligo and EST Anatomy Tool Viewer. This allows users to perform an "electronic northern" for the oligonucleotides using EST frequency data. These web pages are based on the National Cancer Institute's Serial Analysis of Gene Expression (SAGE) Genie (http://cgap.nci.nih.gov/SAGE) with the modification that EST, rather than SAGE data, is being presented. Currently, EST frequency for the oligonucleotides is reported in tabular and graphical formats based on tissue origin of the cDNA library. A separate tool, termed the Highly Expressed Gene Finder page, allows users to select a tissue and generate a list, based on EST frequency, for the most frequently expressed genes. In the future, we will be adding array expression data to these "electronic northern" pages to allow users to mine both data sets for transcript presence and frequency.

SUMMARY

Maize long-oligonucleotide arrays are providing a powerful method for assaying gene expression on a global scale. To date, we have distributed over 900 array sets to researchers in the United States, China, Switzerland, England, Italy, and Mexico. A more complete genome sequence for maize is expected to be available in the near future. The additional sequence combined with empirical results obtained with the first version of the long-oligonucleotide arrays will undoubtedly facilitate the design of an improved set of long-oligonucleotides that will allow a more complete representation of the maize transcriptome.

ACKNOWLEDGEMENTS - We would like to acknowledge Lou Butler for her critical reading of the manuscript and the excellent technical assistance of Tom Watson, Maya Gchachu, and Leukena Cheam in printing the long-oligonucleotide arrays. We would also like to thank Sharon Henne and Nisha Sahay at Operon Biotechnologies for excellent customer service and Sajeev Batra for bioinformatic assistance in designing the long-oligonucleotides

REFERENCES

BENNETT M.D., D.A. LAURIE, 1995 Chromosome size in maize and sorghum using Em serial section reconstructed nuclei. Maydica **40:** 199-204.

BRAZMA A., P. HINGAMP, J. QUACKENBUSH, G. SHERLOCK, P. SPELLMAN, C. STOECKERT, J. AACH, W. ANSORGE, C.A. BALL, H.C. CAUSTON, T. GAASTERLAND, P. GLENISSON, F.C. HOLSTEGE, I.F. KIM, V. MARKOWITZ, J.C. MATESE, H. PARKINSON, A. ROBINSON, U. SARKANS, S. SCHULZE-KREMER, J. STEWART, R. TAYLOR, J. VILO, M. VINGRON, 2001 Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. **29:** 365-371.

BUCKLER E.S.T., J.M. THORNSBERRY, S. KRESOVICH, 2001 Molecular diversity, structure and domestication of grasses. Genet. Res. **77:** 213-218.

CASATI P., V. WALBOT, 2003 Gene expression profiling in response to ultraviolet radiation in maize genotypes with varying flavonoid content. Plant Physiol. **132:** 1739-54.

CONE K.C., M.D. MCMULLEN, I.V. BI, G.L. DAVIS, Y.S. YIM, J.M. GARDINER, M.L. POLACCO, H. SANCHEZ-VILLEDA, Z.W. FANG, S.G. SCHROEDER, S.A. HAVERMANN, J.E. BOWERS, A.H. PATERSON, C.A. SODERLUND, F.W. ENGLER, R.A. WING, E.H. COE, 2002 Genetic, physical, and informatics resources for maize on the road to an integrated map. Plant Physiol. **130:** 1598-1605.

DEYHOLOS M.K., D.W. GALBRAITH, 2001 High-density microarrays for gene expression analysis. Cytometry **43:** 229-238.

FERNANDES J., V. BRENDEL, X. GAI, S. LAL, V.L. CHANDLER, R.P. ELUMALAI, D.W. GALBRAITH, E.A. PIERSON, V. WALBOT, 2002 Comparison of RNA expression profiles based on maize expressed sequence tag frequency analysis and micro-array hybridization. Plant Physiol. **128:** 896-910.

GAUT B.S., J.F. DOEBLEY, 1997 DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. USA **94:** 6809-6814.

MESSING J., A.K. BHARTI, W.M. KARLOWSKI, H. GUNDLACH, H.R. KIM, Y. YU, F. WEI, G. FUKS, C.A. SODERLUND, K.F. MAYER, R.A. WING, 2004 Sequence composition and genome organization of maize. Proc. Natl. Acad. Sci. USA **101:** 14349-54.

MEYERS B.C., S.V. TINGEY, M. MORGANTE, 2001 Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. **11:** 1660-76.

NUWAYSIR E.F., W. HUANG, T.J. ALBERT, J. SINGH, K. NUWAYSIR, A. PITAS, T. RICHMOND, T. GORSKI, J.P. BERG, J. BALLIN, M. MCCORMICK, J. NORTON, T. POLLOCK, T. SUMWALT, L. BUTCHER, D. PORTER, M. MOLLA, C. HALL, F. BLATTNER, M.R. SUSSMAN, R.L. WALLACE, F. CERRINA, R.D. GREEN, 2002 Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res. **12:** 1749-1755.

OSATO N., H. YAMADA, K. SATOH, H. OOKA, M. YAMAMOTO, K. SUZUKI, J. KAWAI, P. CARNINCI, Y. OHTOMO, K. MURAKAMI, K. MATSUBARA, S. KIKUCHI, Y. HAYASHIZAKI, 2003 Antisense transcripts with rice full-length cDNAs. Genome Biol. **5:** R5.

PERTEA G., X. HUANG, F. LIANG, V. ANTONESCU, R. SULTANA, S. KARAMYCHEVA, Y. LEE, J. WHITE, F. CHEUNG, B. PARVIZI, J. TSAI, J. QUACKENBUSH, 2003 TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics **19:** 651-652.

PHIMISTER B., 1999 Going global. Nature Genet. **21:** 1-1.

QUACKENBUSH J., J. CHO, D. LEE, F. LIANG, I. HOLT, S. KARAMYCHEVA, B. PARVIZI, G. PERTEA, R. SULTANA, J. WHITE, 2001 The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res. **29:** 159-164.

SALAMOV A.A., V.V. SOLOVYEV, 2000 Ab initio gene finding in Drosophila genomic DNA. Genome Res. **10:** 516-522.

SCHENA M., D. SHALON, R.W. DAVIS, P.O. BROWN, 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467-470.

SHAROPOVA N., M.D. MCMULLEN, L. SCHULTZ, S. SCHROEDER, H. SANCHEZ-VILLEDA, J. GARDINER, D. BERGSTROM, K. HOUCHINS, S. MELIA-HANCOCK, T. MUSKET, N. DURU, M. POLACCO, K. EDWARDS, T. RUFF, J.C. REGISTER, C. BROUWER, R. THOMPSON, R. VELASCO, E. CHIN, M. LEE, W. WOODMAN-CLIKEMAN, M.J. LONG, E. LISCUM, K. CONE, G. DAVIS, E.H. COE, 2002 Development and mapping of SSR markers for maize. Plant Mol. Biol. **48:** 463-481.

SINGH-GASSON S., R.D. GREEN, Y. YUE, C. NELSON, F. BLATTNER, M.R. SUSSMAN, F. CERRINA, 1999 Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat. Biotechnol. **17:** 974-978.

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, J. QUACKENBUSH, 2003 Enrichment of gene-coding sequences in maize by genome filtration. Science **302:** 2118-20.

XU H., S. ZHANG, D. LIU, C.C. LIANG, 2001 End-labeling of long DNA fragments with biotin and detection of DNA immobilized on magnetic beads. Mol Biotechnol. **17:** 183-185.

YAMADA K., J. LIM, J.M. DALE, H. CHEN, P. SHINN, C.J. PALM, A.M. SOUTHWICK, H.C. WU, C. KIM, M. NGUYEN, P. PHAM, R. CHEUK, G. KARLIN-NEWMANN, S.X. LIU, B. LAM, H. SAKANO, T. WU, G. YU, M. MIRANDA, H.L. QUACH, M. TRIPP, C.H. CHANG, J.M. LEE, M. TORIUMI, M.M. CHAN, C.C. TANG, C.S. ONODERA, J.M. DENG, K. AKIYAMA, Y. ANSARI, T. ARAKAWA, J. BANH, F. BANNO, L. BOWSER, S. BROOKS, P. CARNINCI, Q. CHAO, N. CHOY, A. ENJU, A.D. GOLDSMITH, M. GURJAL, N.F. HANSEN, Y. HAYASHIZAKI, C. JOHNSON-HOPSON, V.W. HSUAN, K. IIDA, M. KARNES, S. KHAN, E. KOESEMA, J. ISHIDA, P.X. JIANG, T. JONES, J. KAWAI, A. KAMIYA, C. MEYERS, M. NAKAJIMA, M. NARUSAKA, M. SEKI, T. SAKURAI, M. SATOU, R. TAMSE, M. VAYSBERG, E.K. WALLENDER, C. WONG, Y. YAMAMURA, S. YUAN, K. SHINOZAKI, R.W. DAVIS, A. THEOLOGIS, J.R. ECKER, 2003 Empirical analysis of transcriptional activity in the Arabidopsis genome. Science **302:** 842-846.

YANG Y.H., S. DUDOIT, P. LUU, D.M. LIN, V. PENG, J. NGAI, T.P. SPEED, 2002 Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucl. Acids Res. **30:** e15.

YU L.X., T.L. SETTER, 2003 Comparative transcriptional profiling of placenta and endosperm in developing maize kernels in response to water deficit. Plant Physiol. **131:** 568-582.