

Characterization of the Maize Genome Array

Wenyong Xu

The maize GeneChip® genome array (SYNG007a520046) contains 82661 probe sets including 45 Affymetrix control probe sets, 58 control probe sets (39 genes) for maize house keeping genes and transgenes, 725 probe sets from GeneBank Full-Length cDNA sequences, and other maize EST assembly sequences. Maize GeneChip design was based on the 87,572 transcripts, including 50,058 with protein similarity and 37,514 unknown. Among 82661 probe sets, 32177 probe sets are with no description. There are three primary sequence sets associated with this chip. These have been placed onto Morpheus (<http://morpheus/blast/blast.html>) for BLAST & SeqRetrieve access. For Affymetrix control genes, each probe set contains 11 or 20 pairs of perfect and mismatch. For maize genes, each probe set contains 14-16 perfect-match probes with feature-size 11 μm and no mismatch probes. Probe sequence selection is based toward the 3'-end of the ORF. The 45 Affymetrix controls are designed to qualify overall hybridization efficiency, monitor the quality of chip manufacture, and provide a built in negative control. 10 maize house-keeping genes are designed as positive control, including actin, tubulin, cyclophilin, glyceraldehyde 3-phosphate dehydrogenase (GAPDH), translation initiation factor, etc. The 715 full-length mRNAs were used as seeds to accelerate computation by pulling out sequences stemming from known transcripts. This study determined the basic quality parameters of the maize genome array, and identified ~1700 contigs in sense direction.

The experiments were designed to answer the following questions:

1. What is background threshold level and what is expression level that can be considered as meaningful measurement of the expression data?
2. How reproducible is the maize genome array for technique replicate?
3. How can one assess the quality of the cRNA preparation?
4. What's the detection sensitivity and dynamic range of the maize genome array?
5. How could one identify the orientation of contigs, which was used for designing probe sets?

Experiment design:

The labelling process was started with 2ug DNA with biotin-dNTP in the presence of random hexamers using DNA Klenow fragment at 37°C for 2 hours and pool tubes' product together. The labelled DNA was hybridized to the array overnight (we use one and half tubes' labelled product for each chip hybridisation. In other word, we use 3ug DNA for each hybridisation.). The array was scanned after post-hybridization wash.

Maize RNA (maize cob-pith sample) was isolated and subjected to cDNA and cRNA synthesis. cRNAs were hybridized to the arrays and data were used to compute reproducibility of the maize genome array. Total RNA was extracted using RNAwiz and purified by RNeasy kit. The maize RNA was subjected to cDNA and cRNA synthesis for RNA hybridization.

Technical replicates of cRNA are prepared for the determination of dynamic range, sensitivity, reproducibility, and background noise using pooled total. Once all cRNAs for technical replicates have been created they are subjected to QC analysis to determine the quality and quantity of preparations. These individual preparations are then pooled and from this pool 17 aliquots of 12.5 μg cRNA are produced for use with the spike dilution experiment. These 17 technical replicates are individually spiked with a cocktail of labelled cRNAs from selected control genes (bioB, bioC, bioD, and cre). This cocktail consists of equimolar concentrations of each control gene's labelled cRNA fragments. A serial dilution of this cocktail across a range of 0 – 1500 pM is used for spiking the technical replicates. A plot of the expression values versus spike concentration will define the saturation level and linear range of resolution. Equal molars of spike controls (including BioB, BioC, BioD and CreX) were mixed and serial dilution of the spike controls were hybridized to the maize genome arrays with pool of maize cRNAs. The range of spike controls used was shown below, and after hybridization to the maize genome array, data were collected and used to compute the dynamic range of the maize genome array.

Spike	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R
BioB (pM)	1500	1000	800	500	400	250	100	40	20	8	4	2	1	0.4	0.2	0.1	0
BioC (pM)	1500	1000	800	500	400	250	100	40	20	8	4	2	1	0.4	0.2	0.1	0
BioD (pM)	1500	1000	800	500	400	250	100	40	20	8	4	2	1	0.4	0.2	0.1	0
Cre (pM)	1500	1000	800	500	400	250	100	40	20	8	4	2	1	0.4	0.2	0.1	0

Results:

Gene detection in maize genome array

10 maize chips were used for maize genomic DNA hybridization (B73 and MO17 plant DNA samples) and 19 chips were used for maize RNA hybridization (Cob-pith RNA sample).

Using custom algorithm, data from 10 genomic DNA hybridization experiments were collected. The analysis results indicated that and 99% of probe sets for DNA hybridisation have signal intensity greater than 12.5. 95% of probe sets have signal greater than 25.

19 RNA hybridization results indicated that 91.07% of probe sets have signal intensity greater than 12.5. About 67.44% of genes were detected and called “present” using SBI2 Algorithm.

cutoff	cob-pith-RNA		B73-DNA		MO17-DNA	
	#probesets (detected)	%	#probesets (detected)	%	#probesets (detected)	%
>12.5	75238	91.070	82146	99.431	82118	99.397
>15	71481	86.522	81765	98.970	81703	98.895
>25	54965	66.531	78547	95.075	78291	94.764
>35	42303	51.204	72332	87.552	71751	86.848
>62	24658	29.847	45073	54.557	44332	53.660

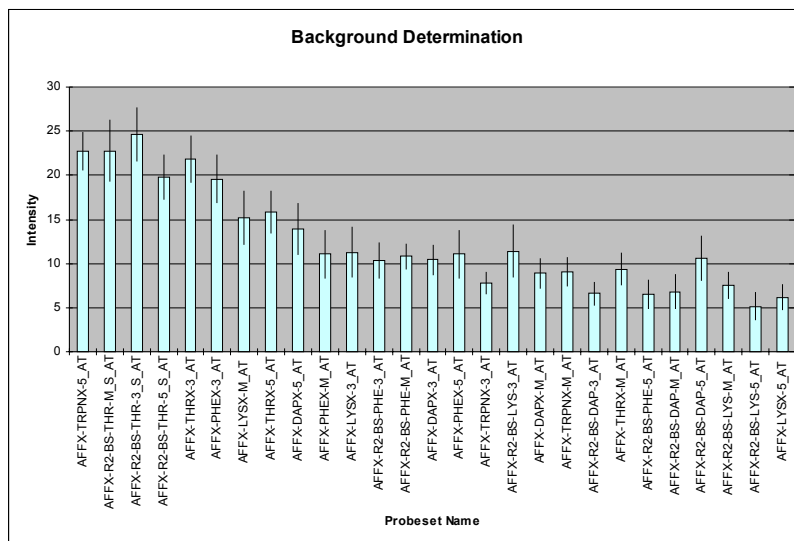
The custom algorithm used for maize GeneChip data analysis:

The custom algorithm SBI2 was employed for maize GeneChip data analysis. This algorithm include following five steps:

1. Outlier detection algorithm determined whether a probe cell is an outlier and removes all probes assigned as outlier. (Note: Outliers are probe cells that are obscured or non-uniform in intensity, for example, probe cells with bright or dark streaks.)
2. Probe level background correction: (1) divide the chip into 16 region; (2) for each region, use the average of lowest 2% probe intensity as the background in this region; (3) for each region, do background correction, set probe intensity as 0 if less than 0.
3. Calculate probe set value: (1) for every probe set, select the PM probes with intensity > 0; (2) the noise level of probe set is the average of probes' pixel standard deviation; (3) calculate the average and CV of the probes based on Huber M-estimator, the average is set as expression value of the probe set.
4. Assign absolute call based on global probe set level background and local probe set noise: the lowest 5% probe set value as global background. If probe set value less than global background, we set "A"; if probe set value more than global background plus 2 times local probe set noise, we set "P"; otherwise, set "M"
5. Scaling: set the average of probe set value as 80.

Background determination for the maize genome array

To determine background noise, the expression values of the built-in Affymetrix negative control probe sets are examined. A histogram analysis is performed to study the range and distribution of values for the negative control probes. Based on this analysis and calculation of mean and median, the approximate level of background noise can be determined. There are 27 negative control probes present in the maize chip. Mean expression levels for these negative control genes based on 19 RNA hybridization chips are 12.48 ± 1.408 .



Mean expression levels for these negative control genes based on 10 DNA hybridization chips are 62.49 +/- 20. 87. We need optimize the DNA labeling and hybridization protocol to increase the hybridization stringency.

Monitor the quality and integrity of cRNAs

Positive control probe sets were designed from 10 maize house-keeping genes are designed to measure the quality of synthesized cRNA. Among them, there are seven genes containing probe sets from the regions of 3'-end, middle and 5'-end. Based on 19 RNA hybridization chips, the ratios of 5', middle probe to 3' probe of these seven genes were calculated and shown in the table.

Gene	probeset	Mean	stdev
Maize actin1	CTRL_MAC1-3_AT	1.00	
	CTRL_MAC1-5_AT	0.63	0.05
	CTRL_MAC1-M_AT	1.31	0.12
MubG1 ubiquitin gene	CTRL_U29159.1-3_AT	1.00	
	CTRL_U29159.1-5_X_AT	0.26	0.02
	CTRL_U29159.1-M_X_AT	0.34	0.03
alpha-tubulin 3.	CTRL_ZMALTUB3-3_AT	1.00	
	CTRL_ZMALTUB3-5_AT	0.42	0.05
	CTRL_ZMALTUB3-M_AT	0.94	0.05
translation initiation factor 5A.	CTRL_ZMTRINF5A-3_AT	1.00	
	CTRL_ZMTRINF5A-5_X_AT	0.53	0.03
	CTRL_ZMTRINF5A-M_S_AT	1.43	0.08
polyubiquitin (MubC5)	CTRL_ZMU29158-3_AT	1.00	
	CTRL_ZMU29158-5_X_AT	0.27	0.02
	CTRL_ZMU29158-M_X_AT	0.43	0.03
glyceraldehyde-3-phosphate dehydrogenase GAPC2	CTRL_ZMU45855-3_AT	1.00	
	CTRL_ZMU45855-5_X_AT	0.47	0.05
	CTRL_ZMU45855-M_AT	1.04	0.05
glyceraldehyde-3-phosphate dehydrogenase GAPC3	CTRL_ZMU45856-3_AT	1.00	
	CTRL_ZMU45856-5_X_AT	0.41	0.04
	CTRL_ZMU45856-M_S_AT	1.04	0.08

The relationship between cDNA length and signal decay were studied. The signal decayed ~54% when the cDNA length from 3'-end is about 1085bp. The longer the cDNA length from 3'-end for each probe is, the less intensity of the probe is.

Gene	5'/3'	cDNA length (bp)
Maize actin1	0.63	942.5
MubG1 ubiquitin gene	0.26	1249
alpha-tubulin 3.	0.42	1249
translation initiation factor 5A.	0.53	672.5
polyubiquitin (MubC5)	0.27	1190.5
glyceraldehyde-3-phosphate dehydrogenase GAPC2	0.47	1153.5
glyceraldehyde-3-phosphate dehydrogenase GAPC3	0.41	1140.5
Average	0.46	1085.36
STD	0.13	209.43
C.V. (%)	29.16	19.30

But for the probe sets in the middle (cDNA length from 3'-end is about 600bp), the intensity decayed seems not very consistent among the different genes and most of “_M_” probe sets show very high intensity.

Gene	M/3'	cDNA length (bp)
Maize actin1	1.31	564.5
MubG1 ubiquitin gene	0.34	750
alpha-tubulin 3.	0.94	750
translation initiation factor 5A.	1.43	403.5
polyubiquitin (MubC5)	0.43	750
glyceraldehyde-3-phosphate dehydrogenase GAPC2	1.04	750
glyceraldehyde-3-phosphate dehydrogenase GAPC3	1.04	750
Average	1.005	617
STD	0.41	137.86
C.V. (%)	40.98	22.34

Identify the orientation of contigs using RNA hybridisation data

About 3544 probe sets were designed from both contig sequences and reverse complement contigs. Among them, there are 1752 probe sets for 1667 contigs with “_C”, and 1792 probe sets for 1705 contigs with “_RC”.

Total number of Contigs with no orientation in maize chip	3373(3544 probesets)
# contigs with intensity (RNA in Cob-pith) \geq 25 detectable for contig orientation (either \geq 25)	1455 (1527 probesets)
undetectable contig orientation(both $<$ 25)	1184
undetectable contig orientation (both \geq 25)	712
Contigs only in one direction	39

Employ the RNA hybridisation data to determine the orientation of the contigs (see summary in table):

- Download the RNA expression data from ZB001 (19 replicate chips for cob-pith RNA expression) for the 3544 probe sets, which were designed, based on 3373 contigs with no orientation information.
- Average the intensity for the probe sets with "_C" and "_RC" contigs.
- Using signal intensity 25 as criteria, 1455 contigs (1527 probe sets) have signal intensity greater than 25. (Note: 25 is the top level of signal intensity for Affymetrix negative control probe sets.
- The intensity of probe sets for contigs either with "_C" or "_RC" is \geq 25, call “**detectable for contig orientation**”; the intensity of probe set for contigs both with "_C" and "_RC" is \geq 25 or $<$ 25, call “**undetectable for contig orientation**”. In sum, 1435 contigs’ orientation may be detectable. 1184 contigs have signal intensity greater than 25 and 712 contigs have signal intensity lower than 25 in both directions. The undetectable contigs may be related to the limited gene expression in maize cob-pith RNA or other biological reasons.
- 39 contigs with “_RC” and do not have matched contigs, so their orientation is “undetectable”.

Reproducibility

Reproducibility is measured by comparison of the replicates. The technical replicates were examined. The replicate experiments allow for the determination of the reproducibility of chip data and provide an estimate of the expected variation due to sampling and processing.

To measure reproducibility of expression profiling data, maize cob-pith cRNA samples were pooled and 19 technical replicates showed the good correlation coefficient. The average of Corr. Coeff. for 171 pairs’ comparisons is 0.988 and C.V. is 0.9%.

	ZB001001	ZB001002	ZB001003	ZB001004	ZB001005	ZB001006	ZB001007	ZB001008	ZB001009	ZB001010	ZB001011	ZB001012	ZB001013	ZB001014	ZB001015	ZB001016	ZB001017	ZB001018	ZB001019
ZB001002	0.997																		
ZB001003	0.989	0.989																	
ZB001004	0.987	0.988	0.997																
ZB001005	0.989	0.989	0.998	0.997															
ZB001006	0.989	0.990	0.998	0.996	0.997														
ZB001007	0.980	0.982	0.983	0.982	0.983	0.988													
ZB001008	0.974	0.976	0.976	0.976	0.977	0.982	0.997												
ZB001009	0.990	0.990	0.998	0.997	0.998	0.998	0.983	0.976											
ZB001010	0.990	0.990	0.998	0.994	0.997	0.997	0.983	0.976	0.998										
ZB001011	0.988	0.989	0.998	0.998	0.998	0.998	0.983	0.977	0.998	0.997									
ZB001012	0.988	0.988	0.998	0.997	0.998	0.996	0.982	0.976	0.997	0.995	0.998								
ZB001013	0.987	0.988	0.998	0.996	0.998	0.998	0.984	0.978	0.998	0.998	0.998	0.997							
ZB001014	0.987	0.988	0.998	0.996	0.998	0.997	0.984	0.979	0.997	0.996	0.998	0.997	0.998						
ZB001015	0.987	0.988	0.996	0.994	0.996	0.998	0.992	0.988	0.995	0.994	0.996	0.995	0.997	0.996					
ZB001016	0.977	0.979	0.980	0.980	0.981	0.986	0.998	0.998	0.980	0.979	0.981	0.980	0.981	0.982	0.991				
ZB001017	0.974	0.976	0.977	0.978	0.979	0.983	0.996	0.998	0.977	0.975	0.978	0.979	0.978	0.980	0.989	0.997			
ZB001018	0.963	0.965	0.966	0.966	0.967	0.973	0.992	0.996	0.966	0.964	0.966	0.967	0.968	0.969	0.981	0.994	0.997		
ZB001019	0.984	0.985	0.989	0.988	0.989	0.993	0.997	0.995	0.989	0.988	0.989	0.989	0.990	0.990	0.996	0.997	0.996	0.991	

Maize leaf labeled DNA samples were pooled and 10 chips (5 technical replicates for B73 and MO17) showed the good correlation coefficient. The average of Corr. Coeff. for 45 pairs' comparisons is 0.976 and C.V. is 0.83%.

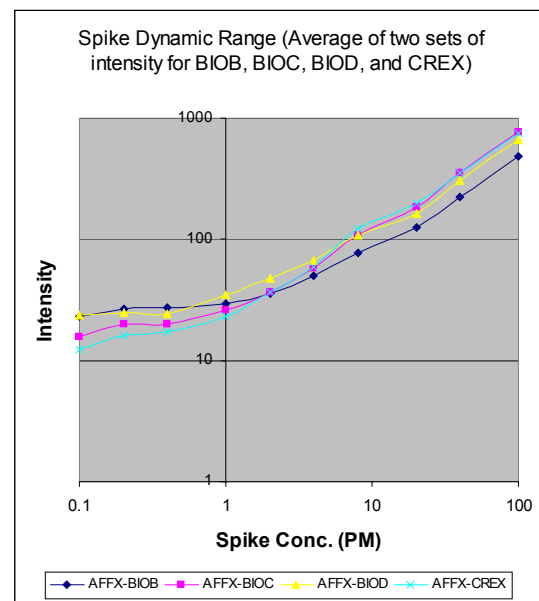
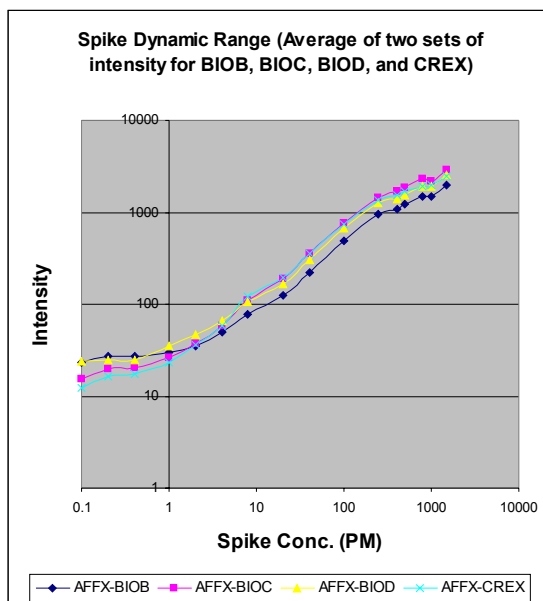
	ZB001020	ZB001021	ZB001022	ZB001023	ZB001024	ZB001025	ZB001026	ZB001027	ZB001028
ZB001021	0.9722951								
ZB001022	0.9566142	0.9786503							
ZB001023	0.9824543	0.9761897	0.9672484						
ZB001024	0.979922	0.9787032	0.9662721	0.9838096					
ZB001025	0.9707484	0.9672736	0.9636	0.9798424	0.979014				
ZB001026	0.9775894	0.9841936	0.9751264	0.9813882	0.984369	0.9797496			
ZB001027	0.9729478	0.9634092	0.9584475	0.9820693	0.9762407	0.9841344	0.975588		
ZB001028	0.976318	0.9851765	0.973821	0.9801505	0.9851183	0.9789895	0.9912671	0.9762802	
ZB001029	0.9861832	0.9674937	0.9504075	0.9810058	0.9770625	0.9726844	0.9778124	0.9779372	0.9776705

A false positive is indicated if a probe is scored quantitatively as changing by at least two fold and the relative hybridised signal is greater than 12.5 (expression levels of Affy-negative control probes). 16 picked replicated hybridization results indicated the false positive rate for the maize genome array was very low with a false positive rate of $(0.052 \pm 0.026) \%$.

Sample ID	Number of FP	Number of probesets	False positive rate (%)
ZB001004 vs. ZB001003	32	82661	0.04
ZB001005 vs. ZB001004	46	82661	0.06
ZB001006 vs. ZB001005	38	82661	0.05
ZB001007 vs. ZB001006	104	82661	0.13
ZB001008 vs. ZB001007	36	82661	0.04
ZB001009 vs. ZB001008	55	82661	0.07
ZB001010 vs. ZB001009	32	82661	0.04
ZB001011 vs. ZB001010	38	82661	0.05
ZB001012 vs. ZB001011	22	82661	0.03
ZB001013 vs. ZB001012	40	82661	0.05
ZB001014 vs. ZB001013	26	82661	0.03
ZB001015 vs. ZB001014	55	82661	0.07
ZB001016 vs. ZB001015	76	82661	0.09
ZB001017 vs. ZB001016	31	82661	0.04
ZB001018 vs. ZB001017	20	82661	0.02
ZB001019 vs. ZB001018	43	82661	0.05

Dynamic Range and Sensitivity

To determine the dynamic range of the maize genome array, an equal molar of spike control mixture including BioB, BioC, BioD and CreX was prepared and serial dilution of spike control mixture was mixed with 12.5 μ g maize cRNAs and hybridized to the maize genome array.



As shown in the figures, the linear dynamic range is between 0.4 pM and 800 and there is greater than 1000 fold linearity for the maize genome array. Using custom algorithm, the sensitivity of detection for the array is around 0.4 pM based on the 3 bacterial spike control genes except BioD which is around 1pM. Using Affymetrix algorithm, the sensitivity of detection will go further to 0.2pM for all four bacterial control genes.

In Conclusions

1. Background level for the maize genome array is less than 25 based on Affy-negative controls in the hybridisation RNA sample on the maize array. For maize DNA hybridization, we need optimize the protocol for hybridization to increase the stringency.
2. Data generated from maize genome array is highly reproducible. The false positive rate among technical replicates of RNA hybridisation is less than 0.1% and the average of Corr. Coeff. is about 0.988. For the technical replicates of genome DNA hybridisation, the average of Corr. Coeff. is about 0.976.
3. The dynamic linear range for the array is greater than 1000 fold and sensitivity of detection for the array is around 0.4 pM based on the at least 3 bacterial spike control genes using different algorithms.
4. In maize RNA hybridisation, about 67.44% of genes were detected and called “present” using SBI2 Algorithm; in genomic DNA hybridisation, 99% probe sets generated signal is greater than 12.5.
6. Based on the ratio of 5’/3’ and 5’/M for seven positive control genes, the intensity decreased to 46% when cDNA length from 3’-end for each probe is 1085bp.
7. There are 3372 contigs (3544 probe sets) with no orientation information in maize chip. Employing the RNA hybridization data, 1455 contigs (1527 probe sets) have signal intensity greater than 25. The orientation for 1435 contigs may be detectable and 1935 contigs could not detectable. A lot of factors need lead to “undetectable”, for example, some of antisense RNA express in tissues rather in cob-pith, the signal intensity cutoff may be not stringent enough or too stringent, etc.